Image Quality Assessment: Utility, Beauty, Appearance

Kwaliteitsbeoordeling van beelden: nut, schoonheid, voorkomen

Ljiljana Platiša

Promotoren: prof. dr. ir. W. Philips, prof. dr. ir. A. Pižurica
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen: Computerwetenschappen

Vakgroep Telecommunicatie en Informatieverwerking
Voorzitter: prof. dr. ir. H. Bruneel
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2013 - 2014

UNIVERSITEIT
GENT

ISBN 978-90-8578-667-2
NUR 958, 954
Wettelijk depot: D/2014/10.500/13

Members of the jury

prof. dr. ir. Rik Van de Walle (Ghent University, chairman)
prof. dr. ir. Heidi Steendam (Ghent University, secretary)
prof. dr. ir. Wilfried Philips (Ghent University, supervisor)
prof. dr. ir. Aleksandra Pižurica (Ghent University, supervisor)
dr. Aldo Badano (U.S. Food and Drug Administration)
prof. dr. Harrison H. Barrett (University of Arizona)
prof. dr. ir. Bart Goossens (Ghent University)
dr. ir. Tom Kimpe (Barco N.V.)
prof. dr. Maximiliaan Martens (Ghent University)
dr. lic. Ewout Vansteenkiste (Ghent University)
prof. dr. Federica Zanca (Catholic University of Leuven)

Affiliations

Research Group for Image Processing and Interpretation (IPI)
Interdisciplinary Institute for Broadband Technology (iMinds)
Department of Telecommunications and Information Processing (TELIN)
Faculty of Engineering and Architecture
Ghent University

Sint-Pietersnieuwstraat 41
B-9000 Ghent
Belgium

Acknowledgments

First and foremost, I thank to my thesis supervisors Prof. Wilfried Philips for setting an example of rigorous but creative scientific thinking, for his questioning attitude and his always constructive feedback, and to Prof. Aleksandra Pižurica for her inspiring enthusiasm for research and science and her much support and encouragement, both scientific and personal. I am grateful to both of them for their dedicated guidance and critical advice as well as for their patience; these have been invaluable in the pursuit of this research and in the preparation of this manuscript.

Also, I thank to Dr. Ewout Vansteenkiste for helping me in my first steps into the world of research and for involving me in the scientific community. A special thank you to Dr. Aldo Badano for the many enriching and fruitful discussions, his frank criticism and insightful comments and suggestions to my work, and his thorough reading and thoughtful editing of our manuscripts.

I would like to thank to the members of my PhD examination committee for the critical reading of this manuscript and the stimulating discussion during the preliminary thesis defense. I believe their questions, comments and suggestions have strengthened the final manuscript.

The benefit of collaboration is indispensable for this work I want to gratefully acknowledge all those who I worked with. I thank to Dr. Leen Van Brantegem for providing great assistance in designing and preparing our human observer experiments for digital pathology data and to Prof. Richard Ducatelle for serving as a "reference" observer in that study as well as for his many thoughtful remarks regarding the principles and challenges of using digital images in veterinary pathology diagnostics. Also providing valuable insight and support to that research were Quentin Besnehard and Dr. Yves Vander Haeghen, as well as Dr. Cédric Marchessoux and Dr. Tom Kimpe. Moreover, Dr. Marchessoux and Dr. Kimpe contributed also to the research and development of model observers through many practical, experience-based discussions and by providing test image data. Further, I want to thank to Dr. Aldo Badano, Dr. Brandon D. Gallas, and Dr. Subok Park for their insightful feedback and valuable suggestions related to the work on model observers for volumetric image data. In particular, I appreciate their comments and suggestions on the data analysis and interpretation of the results and their unreserved assistance and support (especially Brandon's) in publishing the related journal manuscript. Moreover, I am grateful to them for the hospitality during my visit to the Imaging Physics Lab (Division of Imaging and Applied Mathematics, Office of Science and Engineering Labs, CDRH, FDA) in September 2008; it was for me an immensely rewarding and a memorable

experience. Next, I want to thank to Prof. Karel Deblaere M.D. for his expertise, his enthusiasm and dedication in providing feedback to the design and results of our human observer studies of signal detectability as well as for his own participation in the experiments. Also, it was a pleasure collaborating with Nemanja Lukić and Prof. Miodrag Temerinac on real-time implementation of the algorithms for blur estimation proposed in this thesis. Finally, I would like to thank to Prof. Ingrid Daubechies for introducing our group to the intriguing research area of digital artwork analysis and for her inspiration of this work. As a result, several of my IPI colleagues and I are working as part of a magnificent research team involving Prof. Marc de Mey, Prof. Maximiliaan Martens, Dr. Annick Born, Dr. Emile Gezels, Prof. Ann Dooms, and Bruno Cornelis; I credit them all for the pleasant and motivating collaboration. In addition, I thank Saint Bavo cathedral, Lukas - Art in Flanders and the Dierickfonds for their permission to use for my research the digital images of the artwork by Van Eyck which are based on negatives from the late Alfons Dierick photographs (c-04, h-16, 40-15) made available for research purposes to Ghent University. Once more, my sincere thanks to all those who I have worked with.

My sincere thanks extends to all my co-authors for their intellectual insights and wonderful collaboration. Furthermore, I am grateful to those with whom I interacted during the scientific conferences and other meetings. I truly appreciate all your various contributions and insights; they have been invaluable for this dissertation.

My special gratitude goes to all participants in the human observer experiments ran during this dissertation: pathologists Dr. Beatrice Wegge, Dr. Cynthia de Vries, Dr. Koen Chiers, Dr. Leslie Bosseler, Dr. Marjolein van Heerden, and Dr. Sara Van der Heyden; participating students of the Faculty of Veterinary Medicine (school year 2011/2012) and researchers from the TELIN, and especially the volunteers: from the IPI group Jan Aelterman, Danilo Babin, Bart De Spiegeleer, Ivana Despotović, Jonas De Vylder, Koen Douterloigne, Bai Funing, Filip Rooms, Peter Van Hese, Ewout Vansteenkiste, Bart Goossens, Sebastian Gruenwedel, Vedran Jelača, Ljubomir Jovanov (also for developing the camera software), Wenzhi Liao, Benhur Ortiz, Jorge Oswaldo Nino Castaneda, Hiep Quang Luong, and Tijana Ružić, and outside of IPI Biljana Ilić.

It was a great pleasure working and spending time with all my colleagues at IPI-TELIN. I am especially grateful to the colleagues with whom I collaborated on the topic of image quality assessment: Prof. Bart Goossens for his dedicated and thorough review of my first conference submission as well as for the constructive feedback and valuable support around the topic of model observers, Benhur Ortiz for experimenting with my blur algorithms and for bringing his fresh ideas to our "quality team", and Asli Kumcu for her precious collaboration in the investigations involving human observers, her diligent feedback (also to the text of several chapters in this thesis) and our many interesting discussions, and her friendship. For the Dutch translation of my PhD summary, I thank very much to Simon Donné and also acknowledge the kind assistance by Jan Aelterman and Jonas De Vylder, and Claudine Joos (outside of IPI). For all the administrative support, a thank goes to Annette Nevejans, Sylvia Moeneclaey, Alice Verheylesonne, and Patrick Schaillée, as well as to Philippe Serbruyns and Davy Moreels for their exceptional IT support and their much patience and tolerance for my

extensive computer memory requirements. A hearty thanks to my office mate and a very dear friend Tijana Ružić for the many enjoyable chats and laughs and nice moments together. For the pleasant working atmosphere, thank you also to Attila Fesus and Aleksandar Latić and all our former office mates.

My final thanks are to my family. I am deeply grateful to my parents Radoslav and Ljubica for their caring love and unconditional support in all my choices and endeavours, and to my sister Biljana for her enthusiastic encouragement and her unfailing cheering me on, all the way long. Most of all, I thank to my husband Milan for his loving support and his unending patience and understanding throughout my thesis work, and I thank to our son Jovan for giving us the invaluable joy and happiness.

Ljiljana Platiša
Ghent, March 2014

Summary

In today's highly competitive and demanding technology world, digital imaging technologies are continuously being improved and challenged for their excellence in quality and safety (*e.g.* x-ray imaging). Quite impressively, even though current imaging systems achieve high performance in these aspects, there is still much potential for technological and algorithmic advances. It goes without saying that high-end advances in the process of image "production" inevitably bring strong requirements on the related process of image "evaluation" – *image quality assessment*, which is the subject of this thesis.

Although current imaging systems are very advanced, they are inherently imperfect and they produce imperfect images. The quality of an image gets distorted at various stages of the imaging chain, starting from the *image acquisition* process (*e.g.* image blur may be caused by incorrect focus adjustment of a digital camera, or noise may occur in the image due to low radiation dose in medical x-ray imaging), through various *image processing* steps (*e.g.* blocking artifacts introduced in the compression process), up to the *image display* or printing phase (*e.g.* decreased image contrast due to slow temporal response of a liquid crystal display (LCD)). Some of these distortions are rather obvious to the human eye while some remain imperceptible; some of them affect our impression of the image excellence, or influence our ability to perform a certain task which relies on the images, or both.

A very important aspect of image quality assessment is that the method of measuring quality is *relevant* for the application at hand. For instance, let us consider two image viewing devices – a handheld digital photo viewer and a medical display for digital breast mammography. The photo viewer is typically used for viewing personal or other digital camera photos and a user (layperson) expects that their photos (images) "look nice" on the screen (sharp, good contrast and color, minimal noise). Correspondingly, in this example, it seems most appropriate to judge image quality based on the overall impression of image excellence – the technical quality, or the "beauty" of images. In contrast, the medical display is used by medical specialists for a specific task: detecting cancer while visually inspecting breast mammography images. Therefore, in this use case, image quality corresponds to the average success rate for the detection of breast cancer when viewing the images – the "utility" of images. In this dissertation, next to beauty and utility of images, we also explore the quality of "appearance" of objects in the images, *e.g.*, attributes of appearance such as surface smoothness and object symmetry of jewels in digital images of paintings. This kind of analysis could be of great assistance to art history studies of the painterly technique of an artist.

As we have seen, the exact meaning of the quality of images greatly depends on the

application. Correspondingly, this thesis is largely a result of interdisciplinary collaborations which share a common goal of the quantitative characterization of different image quality aspects: quality in the sense of human appreciation of image *beauty* (work with a TV software developer), or quality in terms of image *utility* for a specific purpose (work with medical doctors, a medical display manufacturer, and a related regulatory research center), or even quality of object *appearance* in the images (work with art historical experts). As a complement to image analysis, we also perform multiple psychophysical studies with humans, assessing their rating of the overall quality of images, their performance in a specific (clinical) task, or their judgment of the similarities in appearance of (small spherical) objects in digital images.

For the domain of digital pathology images, a fast-growing area of research, our work makes several important recommendations. Firstly, our results advise against using the psychovisual ratings of IQ collected in a non-task-based experiment (even if the observers are pathology experts), unless, of course, the goal is to assess the beauty of the images. Secondly, the *context* of the experiment with humans should be carefully chosen, an aspect that is not much discussed in literature. According to our results, if a human is asked to judge the quality of an image in an obviously clinical context (involving a specific clinical task) versus a rather technical context (highlighting the technical attributes of quality such as sharpness or noisiness or contrast), the two quality ratings can be quite different. Lastly, our data present a practical illustration of the (possible) disagreement between the two main definitions of image quality assessment – beauty versus utility. Despite being widely discussed, few reports can be found of the experimental data that test the discord.

In this thesis, the problem of assessing image utility is studied for *volumetric* modalities, which are more and more prevalent in medical imaging. In breast imaging, for example, two-dimensional projection mammography images are gradually being augmented by reconstructed tomosynthesis image volumes. At the same time, a wide range of volumetric modalities are already part of standard clinical practice, *e.g.*, MRI brain scans, CT scans of liver or chest, 3D SPECT of bone, and many others. While there is growing evidence of the practical diagnostic benefits of volumetric imaging, techniques for numerical utility-based evaluation of such images are still lacking. In that respect, we propose two novel mathematical models for task-based quality assessment of volumetric images. These so-called *model observers* are inspired by simplifying assumptions about the mechanisms of the human visual system when browsing through a sequence of image slices. Typically, the task of interest is the *detection* of medical signals (lesions) in the image volume. In addition, we review the theoretical background for three other models from the literature to provide a complete overview of the model observers for three-dimensional images. To study the performance of the models, we conduct an experimental comparative analysis for a range of statistically different volumetric images. Furthermore, the dissertation explores and discusses some basic aspects of the potential practical use of the considered model designs.

As a practical application of the model observers, we conduct four studies evaluating the quality of medical image displays. When developing a new medical display, approving it for the market, or making a decision on which clinical display to buy for

the hospital, it is critical to assess the clinical value of the display, *i.e.*, how well it can serve the clinical *task* of interest. Several major contributions of this dissertation are on investigating the effects of *slow response time* of medical LCD monitors in the task of medical signal detection while browsing an image sequence. In practice, clinicians often scroll from one image slice to another faster than the corresponding change in display pixel luminance can be physically completed. Therefore, the displayed image is often a distorted version of the input image. For our experiments, we consider both synthetic and real clinical image data and use state-of-the-art LCD temporal response models to simulate the effects of image browsing. Firstly, our results show a *decrease* in detection performance due to the slow LCD response time, especially at higher browsing rates. Such negative effects have, subsequently, been confirmed by several human observer studies found in the literature. Secondly, we propose a *novel model observer* targeted specifically at the analysis of slow medical LCDs. Conventional implementation of the model restricted the analysis to the luminance values reached at the end of displaying a given image slice (immediately before switching to the next one). Importantly, depending on the details of the luminance changes over time, we find that such models may under- or overestimate detectability. In contrast, our proposed model has access to luminance information sampled over more finely spaced intervals of time, and is shown to be more accurate. Lastly, one model observer study from this dissertation has served as a *preclinical validation* of an actual medical display system entering the market. In addition, those same results were successfully used to pinpoint the characteristic parameters for a subsequent clinical validation study with clinicians.

On the other hand, for the purpose of assessing *attributes* of image beauty, we propose a novel measure of image *blurriness*, which applies to both full-reference and *no-reference* assessment scenario (with and without the ground truth). The proposed method CogACR relies on the average cone ratio (ACR) of the wavelet coefficients which correspond to the strongest edges in the image. CogACR is highly robust to noise and competitive with the state-of-the-art. Furthermore, we address the well-known problem of image quality being dependent on image *content*. In particular, we propose the histogram of ACR values corresponding to dominant edges in the image as an edge-based descriptor of the image content. Moreover, relying on the proposed edge descriptor, a novel measure of image *similarity* is proposed. In contrast to existing similarity measures which depend on the image context, our method quantifies the similarity of the edge-content in the images. As for practical applications, our initial collaborative investigations indicate that the CogACR method can achieve real-time performance even for high-definition (HD) video input. Correspondingly, the method has been integrated into an existing video quality assessment system and tested with a commercially available HD Set Top Box.

Finally, we investigate the images of *artwork* and develop novel methods for quantifying features of appearance of pearls and pearl-like objects in two-dimensional images. Our proposed measures build upon the so-called *spatiogram* representation of the image data, *i.e.*, the image histogram extended with spatial information. Knowing that surface reflectance is among the most notable characteristics of jewels in paintings, it was essential to have spatial information involved in the analysis of pearl

images. First, we propose a method for visualizing the multidimensional spatiogram data; a problem which has not been addressed before. Next, we study a spatiogram similarity measure found in the literature and find good concordance between the measure and the human judgments of similarity between pearl images. At the same time, we point to a major weakness of the existing similarity measure for the analysis of painted objects – the lack of ability to inform about details of dissimilarities. Furthermore, we introduce a (image restoration) method for *matching spatiograms* of different images and use it in our explorative analysis of the dominant factors of the appearance of pearl-like objects. More generally, this technique could be extended to enable virtual style manipulations. Lastly, we propose a set of *novel spatiogram-based measures* which quantify numerically a set of perceptually relevant object features; mainly, the appearance of surface smoothness and several aspects regarding object symmetry. The methods have been evaluated on a range of pearls and beads, both painted and photographed. Overall, the observed agreement between the new measures and the visually observed image features makes the proposed approach a promising candidate for practical use in characterizing *pearls* in paintings. Tentative applications of the proposed techniques and their advancements include assisting art historians in better understanding the differences or similarities between different artists and their ways of painting pearls, as well as artist identification. Beyond the domain of artwork analysis, these kinds of techniques could be extended to several other domains, such as the dermatology imaging where similar techniques could be used, *e.g.*, to characterize the appearance of a lesion skin. Likely, the exact attributes of appearance may need to be redefined but the core idea of the approach would remain the same.

The work in this dissertation yielded a total of 53 scientific publications, consisting of 2 published journal publications (1 as first author), 1 published book chapter (as first author) and 1 book chapter to appear (as co-author), 24 papers published or accepted for publication in the proceedings of international or national conferences (11 as first author), and the remaining 25 abstracts and scientific conference presentations (12 as first author).

Samenvatting

In de huidige competitieve en veeleisende technologische samenleving worden technieken voor digitale beeldverwerking continu kwalitatief verbeterd en op de proef gesteld. Meer nog: ondanks de aanzienlijke prestaties van de huidige methodes is er nog altijd veel potentieel voor zowel technologische als algoritmische verbeteringen. Het spreekt vanzelf dat dergelijke hoogstaande ontwikkelingen op het vlak van beeldvorming uiteindelijk leiden tot strenge eisen voor het evalueren van de afbeeldingen: dit is waar deze thesis zich op toelegt.

Hoewel ze steeds krachtiger en krachtiger worden, lijden methodes voor beeldvorming onder inherente beperkingen zodat de afbeeldingen bijgevolg niet perfect zijn. Een afbeelding verliest kwaliteit tijdens de beeldvorming: hier kan een slecht ingesteld brandpunt van een digitale camera kan leiden tot vage beelden, of ruis kan in het beeld optreden door lage stralingsdosis in de medische x-stralen beeldvorming. In een volgende stap wordt de afbeelding verwerkt: typisch wordt een afbeelding gecomprimeerd wat kan leiden tot blokartefacten. Tenslotte moet het beeld ook nog weergegeven worden, zij het via een scherm of door middel van een afdruk, waarbij het waargenomen beeld kan verschillen van het beeld dat we willen weergeven, bijvoorbeeld door een laag contrast. Sommige van deze verstoringen zijn duidelijk waarneembaar met het menselijk oog terwijl andere nauwelijks zichtbaar zijn; sommige beïnvloeden de waargenomen beeldkwaliteit, andere – of diezelfde – hebben een impact op ons vermogen om een bepaalde taak uit te voeren aan de hand van het beeld.

Het belangrijkste element voor het grootste deel van de toepassingen van beeldverwerking is om over een objectieve en betrouwbare methode te beschikken om de kwaliteit van een gegenereerde afbeelding te kunnen uitdrukken. Bovenal moet de kwaliteitsschatting *relevant* zijn voor de specifieke toepassing. Voor de producent van digitale fotokaders is de belangrijkste duiding voor kwaliteit de algemene indruk van de beelden – de "schoonheid" van de afbeeldingen. Aan de andere kant staat dan bijvoorbeeld een producent van medische beeldschermen. Voor een beeldscherm dat een mammografie moet weergeven zou een voorspelling van de kans om een aanwezige kanker te identificeren op een afbeelding een veel belangrijkere kwaliteitsmaat zijn – de "bruikbaarheid" van de afbeeldingen. Uiteindelijk zullen we het concept van beeldkwaliteit invullen voorbij de conventionele grenzen van schoonheid en bruikbaarheid, om tot een uitdrukking te komen van de kwaliteit van de "weergave" van objecten in de afbeeldingen. Als voorbeeld geven we de verschillende aspecten van juwelen in digitale afbeeldingen van schilderijen; een manier om hun gladheid en symmetrie numeriek uit te drukken kan een grote hulp zijn bij de studie van schildertechnieken van een bepaalde kunstenaar.

Het is duidelijk dat de precieze betekenis van beeldkwaliteit sterk afhangt van

de specifieke toepassing waarin de beelden gebruikt worden. Bijgevolg is deze thesis grotendeels het resultaat van interdisciplinaire samenwerking met verschillende partijen die allen het quantificeren van beeldkwaliteit als doel hebben: zij het de esthetische schoonheid van een afbeelding in samenwerking met de producent van televisietoestellen, de bruikbaarheid van een afbeelding bij een bepaalde medische taak, of zelfs de kwaliteit van het algemene voorkomen van een object in de afbeeldingen van kunstwerken. Naast de beeldanalyse voeren we ook verschillende psychofysische studies uit om de schoonheid, bruikbaarheid of gelijkaardigheid van twee voorwerpen in de ogen van een menselijke waarnemer te staven.

Ons onderzoek geeft verschillende belangrijke aanbevelingen voor het domein van beeldvorming in de pathologie – een van de snelst groeiende velden in het onderzoek rond beeldvorming in het nabije verleden. Eerst en vooral noteren we dat de expertise van de waarnemers onder de loep moet genomen worden wanneer een psychovisueel experiment ontworpen wordt voor de evaluatie van pathologische diagnosestelling. Onze resultaten tonen aan dat het misleidend kan zijn om de psychovisuele respons van leken te gebruiken in een schatting van de bruikbaarheid van afbeeldingen. Daarnaast dient de context van het experiment duidelijk gekozen en uitgedrukt te worden, een aspect dat totnogtoe weinig aandacht kreeg in de literatuur. Wanneer iemand gevraagd wordt zijn indruk te geven omtrent de kwaliteit van eenzelfde afbeelding, enerzijds in een duidelijke klinische context met diagnose als doel en anderzijds in een zuiver technische context met een focus op scherpte en ruis, kan de kwaliteitservaring sterk variëren afhankelijk van het veronderstelde scenario. Tenslotte vormen onze data een praktische illustratie van de (mogelijke) discrepantie tussen de twee voornaamste vormen van kwaliteitsweergave: schoonheid tegenover bruikbaarheid. Hoewel deze tweedracht uitgebreid besproken wordt, zijn er slechts weinig verslagen met experimentele data die bewijs leveren voor een dergelijke opsplitsing.

In de medische beeldverwerking is er een toenemende trend naar het gebruik van ruimtelijke modaliteiten. Bijvoorbeeld in het geval van mammografie worden de tweedimensionale beelden langzaamaan uitgebreid tot gereconstrueerde volumes aan de hand van tomosynthesis. Tegelijkertijd is een breed gamma aan ruimtelijke modaliteiten al deel van de doorsnee klinische behandeling, bijvoorbeeld in het geval van een MRI van de hersenen, CT scans van de lever of de torso, 3D SPECT van botten, ezv. Hoewel er steeds meer bewijs is voor het praktisch nut van dergelijke ruimtelijke beeldvorming bij het diagnoseproces, zijn de technieken om numerieke bruikbaarheidanalyse van de beelden uit te voeren vaak ontoereikend. Op dit vlak stellen we twee nieuwe wiskundige modellen voor om de kwaliteit van ruimtebeelden bij een specifieke taak in te schatten. De zogenoemde *model observers* worden geïnspireerd door eenvoudige veronderstellingen over het menselijke oog bij het bladeren door een opeenvolging van vlakken uit een volumebeeld. Hetgeen ons het meest interesseert in medische beeldvorming is typisch het *detecteren* van medische signalen (letsels) in de afbeeldingen. Verder bespreken we de theoretische achtergrond van een reeks gerelateerde modellen en voeren we experimenten uit om deze methoden te vergelijken aan de hand van statistisch verschillende volumebeelden. Bovendien onderzoekt en bespreekt de dissertatie ook een aantal basisaspecten van de mogelijke praktische toepassing van de overwogen modelleringen.

Als een praktische toepassing van de model observers voeren we vier studies uit om de kwaliteit van medische displays te evalueren. Bij het ontwikkelen van een nieuw medisch display, het goedkeuren voor productie, of het aankopen van displays door een hospitaal is het belangrijk een schatting van de klinische waarde van het display te vormen, in hoeverre het geschikt is voor de taak die uitgevoerd dient te worden. Een belangrijke bijdrage van deze dissertatie is het onderzoek naar de effecten van *(lage) latentie* bij medische displays bij het bladeren door een beeldsequentie. We bekijken zowel artificiële als echte klinische beelden en we gebruiken state-of-the-art modellen voor de temporele respons van een LCD scherm voor het simuleren van de vertragingen bij het bladeren door een reeks afbeeldingen. Bij het bekijken van de modellen stellen we vast dat het integreren van de informatie over de within-frame luminantie in de modellen een belangrijke stap is. Afhankelijk van het gedrag van de luminantie kan het negeren ervan leiden tot een over- of onderschatting van de kans op detectie. Onze experimentele resultaten waarschuwen voor een *daling* van de nauwkeurigheid van de detectie bij een te trage responstijd van de schermen, vooral wanneer de waarnemer snel doorheen de afbeeldingen bladert; dergelijke negatieve effecten werden al bevestigd door verschillende studies met menselijke waarnemers. Bovendien werd een studie van een model observer uit deze dissertatie gebruikt bij de *preklinische validatie* van een medisch display dat de markt betrad. Daarnaast werden diezelfde resultaten met succes gebruikt om de karakteristieke parameters vast te leggen voor de opvolgende klinische validatie.

Met het doel om de schoonheid van een afbeelding uit te drukken zonder een referentiebeeld te vereisen, stellen we een nieuwe metriek voor om de wazigheid van een afbeelding uit te drukken aan de hand van de zogenaamde Average Cone Ratio (ACR) van wavelet coëfficiënten. De voorgestelde methode is bijzonder robuust in de aanwezigheid van ruis en kan wedijveren met de huidige state-of-the-art methodes. Bovendien duiden onze eerste resultaten erop dat de methode erin slaagt een hoge definitie video in real-time te verwerken. Verder bespreken we ook het bekende probleem dat beeldkwaliteit sterk afhankelijk is van de beeldinhoud. We stellen voor het histogram van ACR waarden corresponderend met de dominante randen in de afbeelding te gebruiken als een randgebaseerde descriptor van de beeldinhoud. Aan de hand van deze descriptor stellen we bovendien ook een nieuwe uitdrukking van de gelijkaardigheid van twee beelden voor. In tegenstelling tot de bestaande vergelijkingsmethodes die afhangen van de beeldinhoud quantificeert onze methode de gelijkaardigheid van twee beelden aan de hand van de randen in de afbeeldingen.

Tenslotte onderzoeken we afbeeldingen van kunstwerken en ontwikkelen we methodes om aspecten van het voorkomen van parels en parelachtige voorwerpen in tweedimensionale beelden te beschrijven. De maatstaven die we voorstellen bouwen verder op de zogenaamde spatiogram-representatie van de beeldinhoud. De voorgestelde methodes worden geëvalueerd op een verzameling van parels en kralen, zowel geschilderd als gefotografeerd. De overeenkomst tussen de nieuwe maatstaven en de waargenomen kenmerken zorgt ervoor dat de voorgestelde methode een veelbelovende kandidaat is voor het karakteriseren van *parels* in schilderijen. Een eerste toepassing van de voorgestelde technieken en hun uitbreidingen is bijvoorbeeld het verlenen van hulp aan kunstgeschiedkundigen om de gelijkenissen of verschillen tussen ver-

schillende artiesten beter te begrijpen, of om een artiest te identificeren aan de hand van zijn methode om parels te schilderen. Naast toepassingen bij kunstanalyse kunnen deze technieken ook uitgebreid worden naar andere domeinen, zoals dermatologie (waarbij we het voorkomen van een wonde willen kunnen beschrijven). In deze andere gevallen zullen de specifieke kenmerken waarschijnlijk moeten opnieuw gedefinieerd worden, maar de basisprincipe blijft hetzelfde.

Het onderzoekswerk in deze scriptie leidde tot een totaal van 53 wetenschappelijke publicaties. De lijst bestaat uit 2 publicaties in tijdschriften (1 als eerste auteur), 2 hoofdstukken in gepubliceerde boeken (1 als eerste auteur), 24 papers in de verslagen van (inter)nationale conferenties (11 als eerste auteur), en verder nog 25 abstracts en presentaties op wetenschappelijke conferenties (12 als eerste auteur).

Contents

List of Acronyms

1D	One dimensional
2D	Two dimensional
3D	Three dimensional
ACR	Average cone ratio
AFC	Alternative forced choice
ANOVA	Analysis of variance
APR	Average point ratio
AUC	Area under the receiver operating characteristic curve
BKE	Background known exactly
BKS	Background known statistically
BL	Blur level
CAD	Computer-aided diagnosis
CNB	Correlated noise background
CHO	Channelized Hotelling observer
CIO	Channelized ideal observer
CLB	Clustered lumpy background
CogACR	Center of gravity of HistACR
CT	Computed tomography
CWT	Continuous wavelet transform
DBlur	Defocus blur
DBT	Digital breast tomosynthesis
DDOG	Dense difference-of-Gaussians
DOG	Difference-of-Gaussians
DWT	Discrete wavelet transform
FN	False negative
FOM	Figure of merit
FOV	Field of view
FP	False positive
FR	Full reference
FROC	Free-response receiver operating characteristic
FSM	Film screen mammography
GBlur	Gaussian blur
GUI	Graphical user interface
HD	High definition
HistACR	Histogram of ACR coefficients
HO	Hotelling observer

HVS	Human visual system
IO	Ideal observer
IQ	Image quality
IQA	Image quality assessment
IQR	Interquartile range
JAFROC	Jackknife free-response receiver operating characteristic
JND	Just noticeable difference
LB	Lumpy background
LCD	Liquid crystal display
LG	Laguerre-Gauss
MBlur	Motion blur
MCMC	Markov-chain Monte Carlo
MDS	Multidimensional scaling
MDNOS	Median opinion score
MOS	Mean opinion score
MRI	Magnetic resonance imaging
MRMC	Multiple-reader multiple-case
msCHO	Multi-slice channelized Hotelling observer
MSE	Mean square error
MWW	Mann-Whitney-Wilcoxon
NR	No reference
PC	Percent correct
PDF	Probability density function
PET	Positron emission tomography
PLS	Partial least squares
PSNR	Peak signal-to-noise ratio
REF	Reference image
ROC	Receiver operating characteristic
ROI	Region of interest
RR	Reduced reference
SKE	Signal known exactly
SKS	Signal known statistically
SNR	Signal-to-noise ratio
SPECT	Single-photon emission computed tomography
SROCC	Spearman rank-order correlation coefficient
ssCHO	Single-slice channelized Hotelling observer
SVD	Singular value decomposition
TN	True negative
TP	True positive
umsCHO	Upsampled multi-slice channelized Hotelling observer
vCHO	Volumetric channelized Hotelling observer
WNB	White noise background

1

General introduction

This dissertation researches the problem of evaluating the quality of digital images. In general, image quality (IQ) means different things for different applications. We develop techniques for three main application areas: medical images (how useful the images are for a given clinical task), multimedia images (what is the overall technical excellence of the images, *e.g.* level of image blurriness), and digital images of art paintings (how an object of interest appears in the images, *e.g.* degree of surface smoothness). As a complement to image analysis, we also perform multiple psychophysical studies with humans, either to learn about human "preferences" or to allow comparative analysis of model versus human performance.

1.1 Problem statement

Digital imaging technologies are continuously being improved and challenged for their excellence in quality and safety (*e.g.* x-ray imaging). Quite impressively, even though current imaging systems achieve high performance in these aspects, there is still much potential for technological and algorithmic advances. It goes without saying that high-end advances in the process of image "production" inevitably bring strong requirements on the related process of image "evaluation" – *image quality assessment* (IQA), which is the subject of this thesis.

Imaging systems are inherently imperfect and they produce imperfect images. The quality of an image gets distorted at various stages of the imaging chain, starting from the *image acquisition* process (*e.g.* image blur may be caused by incorrect focus adjustment of a digital camera, or noise may occur in the image due to low radiation dose in medical x-ray imaging), through various *image processing* steps (*e.g.* blocking artifacts introduced in the compression process), up to the *image display* or printing phase (*e.g.* decreased image contrast due to slow temporal response of a liquid crystal display (LCD) in sequence-browsing mode of image viewing). Some of these distortions are rather obvious to the human eye while some remain imperceptible; some of them affect our impression of the image excellence, or influence our ability to perform a

certain task which relies on the images, or both.

A very important aspect of IQA is that the method of measuring IQ is relevant for the *application* at hand. For instance, let us consider two image viewing devices – a handheld digital photo viewer and a medical display for digital breast mammography. The photo viewer is typically used for viewing personal or other digital camera photos and a user (layperson) expects that their photos (images) "look nice" on the screen ("sharp", "good" contrast and color, "minimal" noise). Correspondingly, in this example, it seems most appropriate to judge IQ based on the overall impression of (physical) image excellence – the technical "beauty" of images, also referred to as the *technical* IQ (TechIQ). In contrast, the medical display is used by medical specialists for a specific task: detecting lesions of breast cancer while visually inspecting breast mammography images, detecting lung cancer in chest x-ray images, or differentially diagnosing a skin lesion in (tele)dermatology. Therefore, in the case of a medical display, it seems more appropriate to rate overall IQ based on the average diagnostic performance for a given diagnostic task – the "utility" of images (TaskIQ).

Most commonly, the TaskIQ is studied in the context of medical signal (lesion) detection tasks, such as the aforementioned detection of breast or lung cancer lesions. A counterexample is the case of dermatology images where diagnosis often relies on the *appearance* of the lesion. As an illustration, a critical factor of successful diagnosis of pigmentary disorders in the image are the subtleties between hypopigmentation and depigmentation, or primary versus secondary changes [Philp et al., 2013]. Accordingly, it seems relevant in this case to assess the quality of appearance (ApprIQ) of lesions in the images; the exact attributes of ApprIQ may vary from one pathology to another.

Importantly, this appearance-based image analysis could be of interest also for non-medical applications. One such example is art historical analysis of digital images of art paintings. In that case, instead of evaluating the effects of *image degradation*, it is of interest to assess the effects of *an art painting technique* on the quality of appearance of the painted *objects*. Specifically, in this dissertation, we study the ApprIQ of pearls and pearl-like objects in paintings; *e.g.*, attributes of appearance such as surface smoothness and object symmetry of pearls in digital images of paintings. In order to limit the influence of possible technical image degradations (*e.g.* the presence of image noise might affect the appearance of object's smoothness), our ApprIQ analysis assumes images of high TechIQ (degradation-free images).

Further in the thesis, we refer to TechIQ, TaskIQ, and ApprIQ as "kinds" of IQ.

1.2 Topical outline

For a structured illustration of IQ evolution over a life-cycle of an image, we refer to the formulation of the *IQ circle* proposed by [Engeldrum, 2004]. As shown in Figure 1.1, there are four basic elements of the IQ circle (represented by square blocks) which correspond to different stages in the life-cycle of an image (stages of the imag-

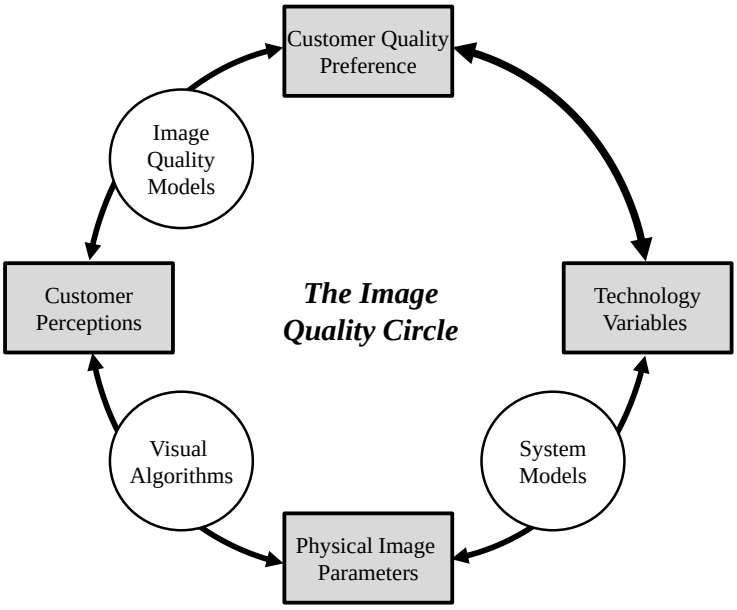


Figure 1.1: The image quality circle [Engeldrum, 2004]

ing chain).[1] Each stage is characterized by a different set of parameters which determine the IQ:

1. *Technology variables*, parameters which are typically optimized during the production of an imaging system, *e.g.*, number of dots per pixel, contrast ratio, response time of an LCD system;

2. *Physical image parameters*, *e.g.*, intensity range, frequency spectrum, Lipschitz regularity of edges (detailed in Chapter 5);

3. *Customer perceptions*, or perceived *attributes* of IQ, *e.g.*, image sharpness, brightness, color fidelity; and

4. *Customer quality preference*, a rating of *overall IQ* according to the user of the image.

Another IQ-related model of interest is known as "the efficacy of diagnostic imaging" proposed by [Fryback and Thornbury, 1991]. It is specialized for diagnostic imaging and represents a hierarchical arrangement of measures of diagnostic imaging

[1] Other aspects of the IQ circle focus on an application-independent approach to IQA (TechIQ of an image) and the "quality preference" criterion; we do not refer to those details.

efficacy (or (cost-)effectiveness). The model assumes six different levels of efficacy, of which the first two are of interest for our research: Level 1, described as the "technical efficacy" of diagnostic imaging, and Level 2, "diagnostic accuracy efficacy". Levels 3 to 6 address the consequences of diagnostic imaging, ranging from whether it produces a change in the diagnostic thinking of the physician, all the way to analysis of the efficiency of use of societal resources to provide medical benefit to society. Given the terminology of this thesis, Level 1 of the [Fryback and Thornbury, 1991] model corresponds to the TechIQ of (medical) images while Level 2 translates to what we call the TaskIQ of medical images.

In terms of the end-value of the product (an imaging system), the most important rating of IQ (TechIQ as well as TaskIQ) is that by the customer (corresponding to Stage 4 of the IQ circle). Typically, IQ is validated by means of a human observer study since the image will be used by humans (and not, for example, by computer aided diagnosis systems). In this thesis, an "observer" takes different forms, as illustrated in Figure 1.2. An observer of IQ can be either a *human* or a mathematical *model*. Among humans, the observers are either *expert users* of the images (*e.g.* experts in image and video processing assessing the technical quality of the images, or medical specialists in radiology assessing the quality of medical x-ray images) or they are *"naive"* to the images (*e.g.* non-medical experts assessing the quality of medical images). On the other hand, model observers may be designed to mimic human ratings of IQ, in which case they are referred to as *anthropomorphic models*, or they could estimate the quality of images from the point of view of *information content* which is especially useful for assessing raw data (*e.g.* in designing and optimizing data acquisition hardware or in developing optimal image reconstruction algorithms) [Barrett and Myers, 2004]. Some examples of the information-based models include TechIQ measures such as mean square error (MSE) or peak signal-to-noise ratio (PSNR) quantifying information loss of a given image relative to the "ideal image" (often referred to as the "reference image") and TaskIQ measures of task performance such as the likelihood ratio that describes the optimal decision/estimation strategy (an "ideal observer") in statistical decision theory. In general, optimal images in terms of information content do not ensure optimality in terms of human performance. Nevertheless, as suggested by [] and further elaborated by [Burgess, 1995a], the ideal observer could be used as a gold standard for the evaluation of human observer performance for a large variety of tasks.

An overview of human as well as model observer studies conducted during the course of this dissertation is presented in Figure 1.3. The studies are grouped according to the kind of IQ being assessed: utility (TaskIQ), beauty (TechIQ), or appearance (ApprIQ). We remark that the models developed in this work are not designed specifically to mimic humans, although many design choices have been motivated by assumptions about the human visual system; details are discussed later in the book. Where possible, model performance was compared against humans and the two agreed about the ranking of IQ for different systems (image parameters). This

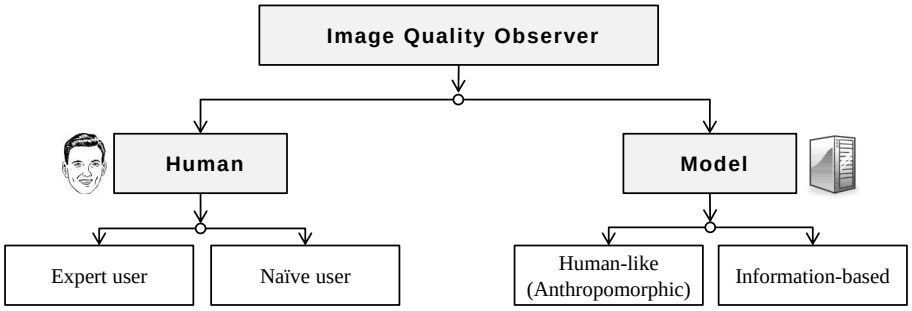


Figure 1.2: Different types of *observers* for the IQA studies.

suggests promising future research on human-like model observers.

1.3 Main contributions and publications

The main contributions resulting from this dissertation are summarized per application area. Moreover, this research greatly benefited from several collaborations and we mention those as well.

TechIQ, TaskIQ – Case study of digital pathology images. For the domain of digital pathology images, a fast-growing area of research, our work makes several important recommendations. Firstly, our results advise against using the psychovisual ratings of IQ collected in a non-task-based experiment (even if the observers are pathology experts). An exception are studies whose primary goal is assessing the TechIQ of the images (in which case technical experts would be preferred over pathology expert observers). Secondly, the *context* of the experiment with humans should be carefully chosen, an aspect that is not much discussed in literature. According to our results, if a human (pathology expert) is asked to judge the quality of an image in an obviously clinical context (involving a specific clinical task) versus a rather technical context (highlighting the technical attributes of quality such as sharpness or noisiness or contrast), the two quality ratings can be quite different. Lastly, our data present a practical illustration of the (possible) disagreement between the two most common approaches to IQA – TechIQ versus TaskIQ. Despite being widely discussed, few reports can be found of the experimental data that test the discord.

This research has been conducted within the framework of the "Color Imaging and Multidimensional Image processing in medical applications" (CIMI) project financially supported by iMinds. The project involved collaboration with multiple academic as well as industrial partners including Dr. Leen Van Brantegem and Prof. Richard Ducatelle (Department of Pathology, Bacteriology and Avian Diseases, Ghent University, Belgium), Quentin Besnehard,

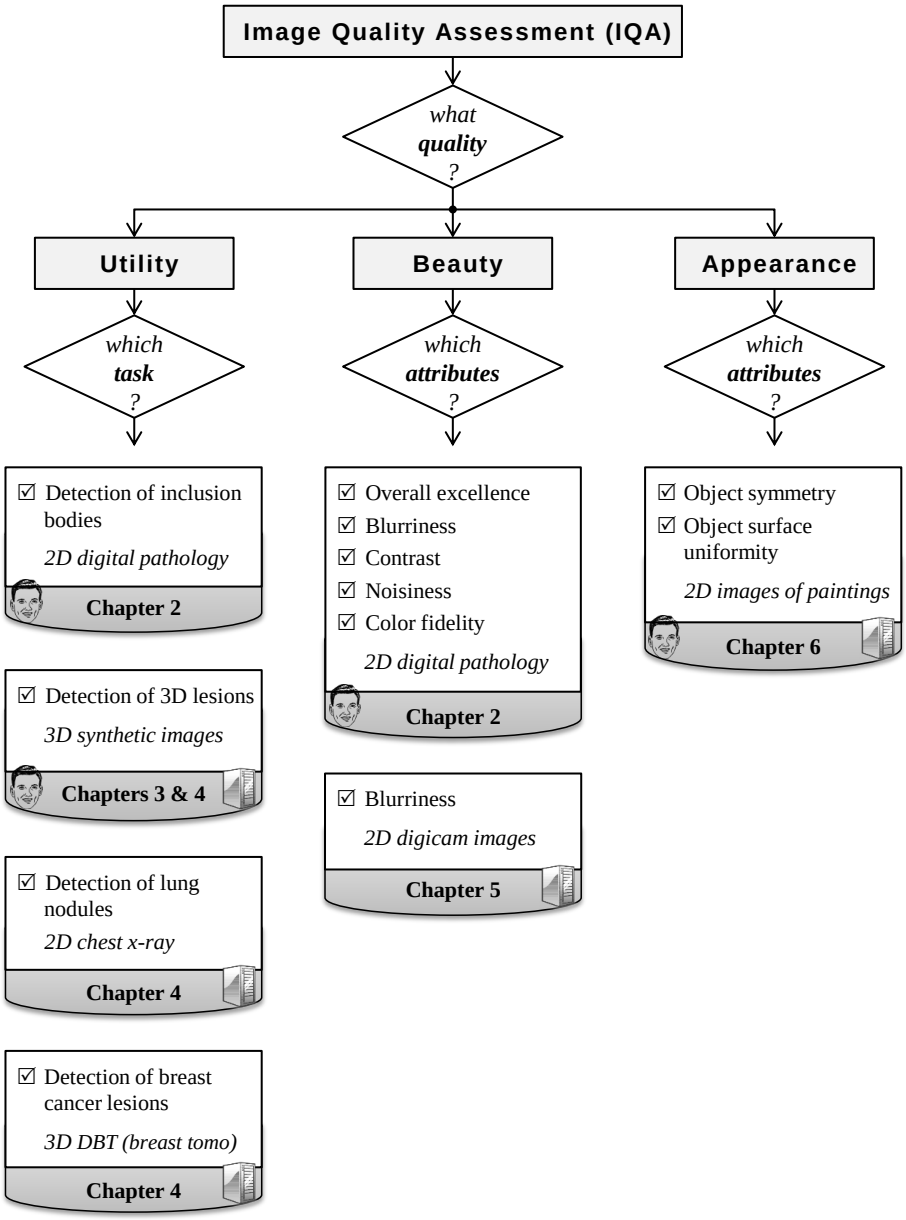


Figure 1.3: Overview of IQA studies in the dissertation.

Dr. Cédric Marchessoux, and Dr. Tom Kimpe (Barco N.V., Belgium), and Dr. Yves Vander Haeghen (Department of ICT, Ghent University Hospital, Belgium). Moreover, all human observer studies in this thesis were conducted in close collaboration with Asli Kumcu (Department of Telecommunications and Information Processing, Ghent University, Belgium).

The thesis work concerning digital pathology images resulted in one conference proceedings [Platiša et al., 2013b] and one conference presentation [Platiša et al., 2013a].

TaskIQ – Models for volumetric detection tasks. While there is growing evidence of the practical diagnostic benefits of *volumetric* imaging, techniques for numerical utility-based evaluation of such images are still lacking. In this thesis, the task of interest is the *detection* of medical signals (lesions) in the image volume. In that respect, we propose two novel mathematical models for task-based quality assessment of volumetric images. These so-called channelized Hotelling observer models, *CHO models*, are inspired by simplifying assumptions about the mechanisms of the human visual system when browsing through a sequence of image slices. In addition, we review the theoretical background for three other CHO models from the literature to provide a complete overview of the model observers for three-dimensional images. To study the performance of the models, we conduct an experimental *comparative analysis* for a range of statistically different volumetric images. Moreover, the dissertation explores and discusses some basic aspects of the *practical use* of the considered model designs.

The results were reported in one journal article [Platiša et al., 2011e], one conference proceedings [Platiša et al., 2009b], and another conference talk [Platiša et al., 2009a].

TaskIQ – Case studies of medical displays. As a practical application of the CHO models, we conduct four studies evaluating the quality of medical image displays. When developing a new medical display, approving it for the market, or making a decision on which clinical display to buy for the hospital, it is critical to assess the clinical value of the display, *i.e.*, how well it can serve the *clinical task* of interest. Several major contributions of this dissertation are on investigating the effects of *slow response time* of medical LCD monitors in the task of medical signal detection while browsing an image sequence. In practice, clinicians often scroll from one image slice to another faster than the corresponding change in display pixel luminance can be physically completed. Therefore, the displayed image is often a distorted version of the input image. For our experiments, we consider both synthetic and real clinical image data and use state-of-the-art LCD temporal response models to simulate the effects of image browsing. Firstly, our results show a *decrease in detection performance* due to the

slow LCD response time, especially at higher browsing rates. Such negative effects have, subsequently, been confirmed by several human observer studies found in the literature. Secondly, we propose a *novel CHO model* targeted specifically at the analysis of slow medical LCDs. Conventional implementation of the model restricted the analysis to the luminance values reached at the end of displaying a given image slice (immediately before switching to the next one). Importantly, depending on the details of the luminance changes over time, we find that such models may under- or overestimate signal detectability. In contrast, our proposed model has access to luminance information sampled over more finely spaced intervals of time, and is shown to be more accurate. Lastly, one model observer study from this dissertation has served as a *preclinical validation* of an actual medical display system entering the market. In addition, those same results were successfully used to pinpoint the characteristic parameters for a subsequent clinical validation study with clinicians.

Next to the model observer studies, we conduct a series of human observer experiments in order to assemble data about human performance for different levels of task difficulty. We compare single-slice versus multi-slice sequence-browsing mode of image viewing. These results aim at guiding future work towards designing a human-like model observer for volumetric image data.

The research related to the model observers and their use for evaluating medical displays has been performed within the framework of the "Medical Virtual Imaging Chain" (MEVIC) project financially supported by iMinds. The project involved collaboration with multiple academic and industrial partners including Dr. Cédric Marchessoux and Dr. Tom Kimpe (Barco N.V., Belgium). In addition, we closely collaborated on these topics with Dr. Aldo Badano, Dr. Brandon D. Gallas, and Dr. Subok Park (U.S. Food and Drug Administration, USA), Prof. Karel Deblaere M.D. (Department of Neuroradiology, Ghent University Hospital, Belgium), Prof. Bart Goossens and Dr. Ewout Vansteenkiste (Department of Telecommunications and Information Processing, Ghent University, Belgium).

This work resulted in six conference proceedings [Platiša et al., 2009d, Platiša et al., 2010c, Platiša et al., 2011g, Platiša et al., 2011h, Platiša et al., 2012c, Kumcu et al., 2012b] and four abstracts and conference presentations [Platiša, 2008, Platiša et al., 2010b, Platiša et al., 2011f, Kumcu et al., 2011c]. A journal article discussing the human observer study of single-slice versus multi-slice image viewing is in preparation [Platiša et al., 2014b].

TechIQ – Models for image blur evaluation. On the other hand, for the purpose of assessing the *attributes* of TechIQ, we propose a novel *no-reference* measure of image *blurriness* based on the average cone ratio (ACR) of wavelet coefficients. The proposed CogACR method is highly robust to noise and competitive with the state-of-the-art. Furthermore, we address the well-known problem of IQ

being dependent on image *content* and propose a novel *edge-based descriptor* of the image content. In addition, relying on the proposed descriptor, a novel measure of *image similarity* is proposed. In contrast to existing similarity measures which depend on the image context, our method quantifies the similarity of the edge-content in the images. This work led to two conference proceedings [Ilić et al., 2009, Platiša et al., 2011j] and two conference talks [Platiša et al., 2010d, Platiša et al., 2011i]. A journal paper is in preparation [Platiša and Pižurica, 2014].

Another conference proceedings is a result of collaboration with Nemanja Lukić and Prof. Miodrag Temerinac (Department for Computing and Control Engineering, Novi Sad University, Serbia) [Lukić et al., 2010]. The proposed CogACR measure has been implemented on a commercially available processor to achieve real-time performance for high-definition (HD) video input. Moreover, with this implementation, the CogACR method has been incorporated in an existing video quality assessment platform and tested with a commercially available HD Set Top Box.

Also, the proposed CogACR method has been successfully used as a tool for video blur estimation in the context of two consortium projects, both financially supported by iMinds: the "Telesurgery" project, which assessed the quality of laparoscopic surgery videos, and the ongoing "Ultra Wide Context Aware Imaging" (PANORAMA) project evaluating the quality of x-ray coronary angiographic image sequences. That work was coordinated by Asli Kumcu (Department of Telecommunications and Information Processing, Ghent University, Belgium).

ApprIQ – Characterizing pearls in art paintings. Finally, we investigate the images of *artwork* and develop novel methods for quantifying attributes of appearance of pearls and pearl-like objects in two-dimensional images. Our proposed measures build upon the so-called *spatiogram* representation of the image data, *i.e.*, the image histogram extended with spatial information. First, we propose a method for visualizing the multidimensional spatiogram data; the problem which has not been addressed before. Next, we study a spatiogram similarity measure suggested by the literature and find a good concordance between the measure and the human judgments of similarity between pearl images. At the same time, we point to a major weakness of the existing similarity measure for the analysis of painted objects – the lack of ability to inform about details (reasons) of detected dissimilarities. Furthermore, we introduce a method for *matching spatiograms* of different images and use it as a tool in our explorative analysis of the dominant factors of the appearance of pearl-like objects. Lastly, we propose a set of *novel spatiogram-based measures* which quantify numerically the appearance of surface smoothness and several attributes regarding object symmetry. The methods have been evaluated on images of painted as

well as of real pearls, and the results suggest significant potential for the new measures to be used as a tool in art historical analysis of pearls in paintings.

We became involved in this research on the initiative of Prof. Ingrid Daubechies (Mathematics Department, Duke University, USA) who put us in contact with Prof. Marc de Mey (Royal Flemish Academy of Belgium for Science and the Arts (KVAB), Belgium) and Prof. Maximiliaan Martens, Dr. Annick Born, and Emile Gezels (Department of Art, Music and Theatre Sciences, Ghent University, Belgium). Together with Prof. Ann Dooms and Bruno Cornelis (Department of Electronics and Informatics, Free University of Brussels, Belgium), we collaborated on developing image processing and analysis tools for art investigation. Our primary focus was the world-famous 15th-century polyptych *Ghent Altarpiece* ("Het Lam Gods" in Dutch) located in the Saint Bavo Cathedral in Ghent. This research coincides with the ongoing five year restoration project of the masterpiece.

The results concerning pearl analysis have been published in two book chapters [Platiša et al., 2012b, Pižurica et al., 2013] and one conference proceedings [Platiša et al., 2011a], and presented in three conference talks [Platiša et al., 2010a, Platiša et al., 2011b, Platiša et al., 2012a]. A journal article is in preparation [Platiša et al., 2014a]. Moreover, this research has been presented to a wider non-technical audience in the form of several talks, newspaper articles, and press releases (listed at the end of Chapter 6).

The work in this dissertation yielded a total of 53 scientific publications, consisting of 2 published journal publications (1 as first author), 1 published book chapter (as first author) and 1 book chapter to appear (as co-author), 24 papers published or accepted for publication in the proceedings of international or national conferences (11 as first author), and the remaining 25 abstracts and scientific conference presentations (12 as first author). Below we include a selection of representative publications which result from the work of this dissertation; a complete list is provided in Appendix A.

- L. Platiša, B. Goossens, E. Vansteenkiste, S. Park, B. D. Gallas, A. Badano, and W. Philips, "Channelized Hotelling observers for the assessment of volumetric imaging data sets," *J. Opt. Soc. Am. A*, vol. 28, pp. 1145–1163, Jun 2011.

- L. Platiša, C. Marchessoux, T. Kimpe, E. Vansteenkiste, A. Badano, and W. Philips, "Channelized Hotelling observers for signal detection in stack-mode reading of volumetric images on medical displays with slow response time," *in Proc. IEEE Medical Imaging Conference*, Oct 23-29, 2011, Valencia, Spain, MIC9.S–292.

- L. Platiša, C. Marchessoux, B. Goossens, and W. Philips, "Performance evaluation of medical LCD displays using 3D channelized Hotelling observers," *in Proc. SPIE Medical Imaging*, Feb 12-17, 2011, Lake Buena Vista, Florida, USA, vol. 7966, pp. 79660T.

- L. Platiša, B. Cornelis, T. Ružić, A. Pižurica, A., Dooms, M. Martens, M. De Mey, and I. Daubechies, chapter *"Spatiogram features to characterize pearls and beads and other small ball-shaped objects in paintings,"* in "Vision and material: interaction between art and science in Jan Van Eyck's time", pp. 315-329, KVAB PRESS, 2012.

- L. Platiša, B. Cornelis, T. Ružić, A. Pižurica, A., Dooms, M. Martens, M. De Mey, and I. Daubechies, "Spatiogram features to characterize pearls in paintings," *in Proc. IEEE International Conference on Image Processing (ICIP)*, Sep 11-14, 2011, Brussels, Belgium, pp. 801-804.

- L. Platiša, L. Van Brantegem, A. Kumcu, C. Marchessoux, E. Vansteenkiste, and W. Philips, "Effects of common image manipulations on diagnostic performance in digital pathology human study," *Medical Image Perception Society Conference XV*, Aug 14-16, 2013, Washington, DC, USA.

- L. Platiša, L. Van Brantegem, Y. Vander Haeghen, C. Marchessoux, E. Vansteenkiste, and W. Philips, "Psycho-visual evaluation of image quality attributes in digital pathology slides viewed on a medical color LCD display," *in Proc. SPIE Medical Imaging*, Feb 9-14, 2013, Orlando, Florida, USA, vol. 8676, pp. 86760J.

- L. Platiša, A. Pižurica, E. Vansteenkiste and W. Philips, "No-reference blur estimation based on the average cone ratio in the wavelet domain," *in Proc. SPIE Electronic Imaging, Multimedia Content Access: Algorithms and Systems V*, Jan 23-27, 2011, San Francisco, USA, vol. 7881B.

- N. Lukić, L. Platiša, A. Pižurica, W. Philips, and M. Temerinac, "Real-time wavelet based blur estimation on cell BE platform," *in Proc. SPIE Conference on Wavelet Applications in Industrial Processing VII*, Jan 17-21, 2010, San José, CA, USA, vol. 7535, pp. 75350C.

1.4 Organization of the dissertation

After the introductory overview in the current chapter, we continue in Chapter 2 with a short summary of the human observer experiments conducted in the course of this dissertation. Using as a practical example the results of our two human observer studies of digital pathology images, we discuss in detail the differences between TechIQ and TaskIQ. Moreover, we make several methodological remarks concerning the design of human observer experiments, the selection of the observers (human subjects), and the related training process.

Chapter 3 addresses the problem of assessing TaskIQ of images. First, we review the state-of-the-art of the mathematical models for task-based IQA of medical images. Typically, the task of interest is the detection of medical signals (lesions). Next, we

present five CHO models which appear to be promising candidates for the treatment of volumetric image data. Two of these models are novel, inspired by simplifying assumptions about the human visual system. Subsequently, we conduct a range of experiments to assess the performance of the models. In particular, four different categories of synthesized images are considered, either with Gaussian statistics or not. The results are discussed in the form of comparative analysis of the models on the basis of different parameters, including the parameters of the image content (difficulty of the task) and the size of the training data set. Finally, we present some practical considerations regarding potential applications for the considered model observer designs.

Chapter 4 demonstrates a practical way of using the models described in Chapter 3. Specifically, we present four model observer studies which evaluate the TaskIQ of either real clinical or synthetic images with regards to the possible effects caused by the display of images. An overview of the considered tasks and images is presented in Figure 1.3. Moreover, we describe a human observer study in which imaging experts examined synthetic images, planar or volumetric data. The observers' task was to make the classification decision – the signal is present, or the signal is absent in the image. The discussion highlights some important aspects of planar (single-slice) versus sequence-browsing (multi-slice) image viewing.

In Chapter 5, we shift the focus to the problem of assessing TechIQ of images. Specifically, we study the effects of image blur, the most common image distortion next to image noise. We first introduce the model of image blur and briefly review the basic principles of multiscale image analysis. Subsequently, we introduce a novel measure of image blurriness which relies on the ability of the wavelet transform to characterize edges in the image. Furthermore, we formulate a novel edge descriptor and explain how it can be applied to the problem of edge-based image matching. For the purpose of comparative performance analysis, we provide also a review of existing techniques for no-reference assessment of image blur. The performance of the proposed techniques is extensively tested and evaluated with respect to a range of parameters, including image content (three public databases are considered), level and type of image blur (Gaussian, defocus, and motion), as well as the presence of varying levels of image noise.

The last topical chapter in this book is Chapter 6 which takes a somewhat unconventional approach to IQA by focusing on the appearance of objects in images. Specifically, we study the attributes of appearance of pearls and pearl-like objects in digital images of paintings. We first introduce the concept of image spatiogram (an extension of image histogram) and investigate the ability of an existing spatiogram similarity measure to quantify overall similarity between pearl images. Then, we introduce a novel method for spatiogram matching and use it in our explorative analysis of the dominant factors of appearance of pearls in the images. The major part of the chapter defines four novel spatiogram-based measures which quantify appearance of surface smoothness as well as several attributes of object symmetry. We test the per-

formance of our proposed techniques for a range of pearls and beads, both painted and photographed. Moreover, we conduct a human observer study to collect human ratings of pearl similarity and pearl appearance and discuss the results in comparison to the performance of our proposed measures.

The book ends with Chapter 7 where we review the main conclusions of the dissertation and draw some inferences from these conclusions, as well as suggest possible directions for future research.

2

Beauty versus utility. Human observer experiments

This chapter illustrates the two main strategies for image quality assessment (IQA): technical (TechIQ) versus task-based (TaskIQ). We conduct two studies with human observers to evaluate the effects of some common image artifacts and image manipulations in the context of digital pathology systems. One study examines the TechIQ and the other study the TaskIQ of the images. Firstly, the collected data is used to facilitate the discussion about the agreement of the TechIQ and the TaskIQ. The motivation for this comparative investigation is in the fact that, on the one hand, the TechIQ experiments are considerably faster and easier to prepare, and often also to conduct, while on the other hand, the TaskIQ experiments are more informative of the actual usefulness of images for the targeted application (*e.g.* the clinical task of detecting lesions). Therefore, it is of interest to determine whether the TechIQ of images can accurately predict the TaskIQ of images. Secondly, in the context of the TechIQ study, we examine the consequences of replacing expert observers (pathologists) by naive people (students) or experts in the imaging field (researchers in image processing), and discuss arising issues. These investigations should indicate if replacing expert pathologists by, for example, imaging experts (who are more readily available) would have an effect on the estimated level of the TechIQ, *i.e.*, on the experimental findings.

2.1 Introduction

The first questions that come with any quality assessment are "what does the quality mean?", "what kind of image quality we want to assess?", "which criterion should we use?". The issue of the appropriate approach to IQA has been much discussed in different fields of both science and industry. Specifically, the argument is between the "beauty" versus the "utility" approach, *i.e.*, should we praise images for how good they look (subjective preference in the technical sense) or should we look at how well the images serve their purpose. The former approach is often referred to as *technical* IQA

(assessment of TechIQ) and the latter one as *task-based* IQA (assessment of TaskIQ).

In this chapter, we focus on IQA by human observers. In general, human rating of TechIQ is based on a very subjective criterion of the *overall impression of quality* for which the thresholds may vary considerably from one individual to another. One way to overcome this inherent subjectivity of the overall quality criteria is to judge some *specific image features* (purely technical such as contrast, spatial resolution, noise, sharpness; or application specific, for example "no skinfolds seen" or "visually sharp reproduction of all vessels" as in film screen mammography (FSM) systems [EUCommission, 1996a]). We refer to the aforementioned two variations of the TechIQ approach as *overall* TechIQ and *feature-based* TechIQ, respectively. In the medical field, for example, the feature-based TechIQ approach is still widely used (although the trends might be changing, as we will discuss shortly) and there are even formal guidelines for this approach covering different imaging modalities. Some examples include radiographic imaging for conventional diagnostic examinations (*e.g.* chest, skull, urinary tract, breast) [EUCommission, 1996a, EUCommission, 1996b], conventional FSM [A.C.R.Committee, 1999] and its successor digital mammography including both computed radiography (CR) and digital radiography (DR) mammogram systems [Williams et al., 2007, Guidelines, 2012, Kanal et al., 2013], as well as computed tomography (CT) [Bongartz et al., 2004]. Obviously, the approach of feature-based TechIQ is applicable to many other imaging domains besides medical, among which are geology, archeology, astronomy, and biology.

As an alternative to the TechIQ approach, we could assess the TaskIQ of images, *i.e.*, how useful images are for a specific task for which they are used. Example tasks include: detecting lesions of breast cancer (for breast mammography images), detecting lung cancer (for chest computed tomography (CT) images), and detecting lesions of multiple sclerosis (for magnetic resonance images (MRI) of brain). Undoubtedly, we expect variability between performances of individuals in the latter approach too, due to variations in experience, age, instructions provided, and multiple other reasons. Nevertheless, because the TaskIQ is measured indirectly by measuring the success of humans in a particular task (rather than merely a preference), the criterion itself - the *level of performance* in the task - is not subjective.

In the case of human observer studies, besides the nature of the criteria for IQA, there are also other considerations to be taken into account when deciding on the preferred approach. One important aspect is certainly the required expertise of the observers. While the TechIQ approach may not necessarily require experts in the field (*e.g.* medical professionals in the case of medical image studies), they are often indispensable under the TaskIQ paradigm. This is simply because of the high specialization of the observer's task (*e.g.* detection of lung nodules in chest CT scans) which requires adequate knowledge and expertise. As a consequence, the TaskIQ observer studies may take much more money as well as time compared to the studies of TechIQ simply because the expert's time is more expensive and less readily available than that of a naive observer. Moreover, human observer studies often require carefully selected

images with "nontrivial" task parameters, *e.g.*, images with subtle (rather than obvious) lesions. This is because, today, we are often evaluating advanced systems of high (rather than low) quality which all perform well for easy tasks. Consequently, there is little value in running experiments for such tasks. On the contrary, it is of interest to assess task performance for difficult tasks, for which we expect most benefit from the modern imaging systems. At the same time, as discussed by [Burgess, 1995b], the images should be chosen in the correct range of human observers' performance in order to ensure the lowest coefficient of variation due to sampling error (limited number of trials). For example, [Burgess, 1995b] found that the 2 alternative forced choice (2AFC) experiments are best done with the proportion of correct responses between 0.85 and 0.95. Yet another important requirement is to know the "ground truth" (gold standard) for the images, to allow for human performance analysis. Thus, all considered, the time and effort required for the preparation of the test images are not to be neglected [Zanca et al., 2012].

Obviously, one crucial aspect of the dispute between the TechIQ versus the TaskIQ approach to the assessment of images is the agreement between their findings. Namely, suppose that we adopt the TechIQ approach to evaluate two different imaging systems A and B and we find that system A is better than system B. The question is, if we then switch to the TaskIQ approach, would we end up with the same conclusion? If yes, then it would be of no consequence which approach we take, TechIQ or TaskIQ. However, if the two approaches were to disagree, then well-grounded scientific arguments need to be put forward for the preferred approach.

At this moment, the scientific community has made strong recommendations for one or the other approach only for a very limited range of applications. One of those is the field of medical IQA for which it is strongly recommended to use the TaskIQ approach [Barrett and Myers, 2004]. Nevertheless, as evidenced by the recent review of mammography clinical evaluation methods in research studies by [Li et al., 2010], there is still a considerable variation among the methods being used in practice. One important reason is the lack of scientific consensus on the recommended methodology for IQA of various medical images. Note, for example, the two aforementioned guidelines for the evaluation of mammography images, one developed by the European Commission [EUCommission, 1996a] and the other one by American College of Radiology [A.C.R.Committee, 1999]. Another possible reason could be the lack of evidence about the (dis)agreement between the different approaches to IQA, *e.g.*, between the TechIQ and the TaskIQ paradigms. To date, only a very few studies have looked into the association between technical quality and diagnostic performance of medical images. Among those, [Taplin et al., 2002] found robust associations between detection of cancers and proper breast positioning (a parameter of feature-based TechIQ) but little or no link between detection and breast compression, contrast, exposure, noise, sharpness, artifacts, and overall quality. The authors note, however, that it might be important to consider more sensitive scales for the parameters of quality before drawing any conclusions. Also related, though with converse

findings, is the study by [Jiang et al., 2007] which compared three reconstruction algorithms for parallel MRI of a fresh bovine liver. The authors found a strong influence of reconstruction algorithm on IQ, both according to the human performance in tumor detection (TaskIQ) and according to a perceptual difference model (TechIQ). While it is not unexpected that the findings differ across different anatomies (human breast versus bovine liver) and different imaging technologies (mammography versus MRI), it is clear that the relationship between TechIQ and TaskIQ is highly non-trivial and requires much further investigation.

In this chapter, we report about two human observer studies for assessing the quality of digital pathology images – a TechIQ and a TaskIQ study. It is important to emphasize that the studies are preliminary and aimed at guiding future more extensive and scientifically rigorous research of the topic. In particular, we are interested in identifying which imaging effects are the dominant factors of IQ (such that later we could study them in more detail) and which IQA concept is preferred and why (so that future studies would follow that approach). Given the preliminary character of the research, the size of the two reported human observer studies is limited (in particular, few test images per condition as detailed in Section 2.3) and prevents from making any strong conclusions. Rather, the results will be used as a discussion aid for some important issues concerning the evaluation of medical images.

Firstly, we assess the influence of some common artifacts of the image digitization and some common image manipulations for digital pathology systems. We do this by means of both the TechIQ and the TaskIQ approach. Secondly, we evaluate the association between the perceived IQ (TechIQ) and the diagnostic value of the images (TaskIQ). By doing this, we hope to provide useful insight into the question of concordance between the TechIQ and the TaskIQ concepts which, as we discussed, is still under-researched. Thirdly, in the case of TechIQ, we examine the consequences of replacing expert observers (pathologists) by naive people (students) or experts in the imaging field (researchers in image processing), and discuss arising issues.

This research has been conducted within the framework of the "Color Imaging and Multidimensional Image processing in medical applications" (CIMI) project financially supported by iMinds. The project involved collaboration with multiple academic as well as industrial partners including Dr. Leen Van Brantegem and Prof. Richard Ducatelle (Department of Pathology, Bacteriology and Avian Diseases, Ghent University, Belgium), Quentin Besnehard, Dr. Cédric Marchessoux, and Dr. Tom Kimpe (Barco N.V., Belgium), and Dr. Yves Vander Haeghen (Department of ICT, Ghent University Hospital, Belgium). Moreover, all human observer studies in this thesis were conducted in close collaboration with Asli Kumcu (Department of Telecommunications and Information Processing, Ghent University, Belgium).

The chapter continues in Section 2.2 with a brief overview of the six observer studies conducted within this dissertation, two of which focused on digital imaging for diagnostic veterinary pathology. We also include a consideration of the current state of research into the problem of IQA for digital pathology. In Section 2.3 we describe the

test images. The two observer studies for pathology images, the TechIQ study and the TaskIQ study, are detailed respectively in Section 2.4 and Section 2.5. The outcomes of the two IQA approaches, TechIQ versus TaskIQ, are contrasted in Section 2.6. Finally, Section 5.9 ends this chapter with a summary of the main conclusions arising from the two studies.

2.2 Psychophysical experiments for IQA

The most straight-forward way to assess the quality of images in the way that humans would do it is through human observer studies. This is especially of interest in the cases where numerical methods for IQA are non-existing or not mature enough, such as those studied within this dissertation. The data collected in observer studies is commonly referred to as human data. Depending on the type of IQA approach, the human data can be either the actual IQ ratings (direct responses) or the various relevant indicators of human performance for a given task (indirect responses, for example, in a lesion detection task, the observer's decision about the lesion being present or not present in the image). Importantly, given the limited sample size of both observers and images in such experiments, the collected human data must be examined through statistical analysis before any inferences or conclusions can be made. Further discussion of the basic concepts (human and model) observer studies can be found in Section 4.1.1.

2.2.1 Overview of human observer studies during the dissertation

In the course of this dissertation, multiple observer studies have been conducted for the purpose of collecting human responses (either direct or indirect) about the quality of images. In Figure 2.1, we give an illustration of the different studies. Some more details about the observers and images from each study are provided in Table 2.1. The studies cover a range of experimental designs, thus permitting a more complete view of the IQA issues (some of which will be discussed in this book). The differentiating aspects between studies include the following:

- IQA approach (TechIQ or TaskIQ)
- application domain (laparoscopic surgery, digital pathology, television broadcast, synthetic medical images);
- profile and number of observers (see Table 2.1);
- number of the experimental trials per observer (see Table 2.1);
- number of image stimuli per trial and associated observer's task (single-stimulus in which a single image is assessed at a time, or double-stimulus in which the two images are assessed comparatively);

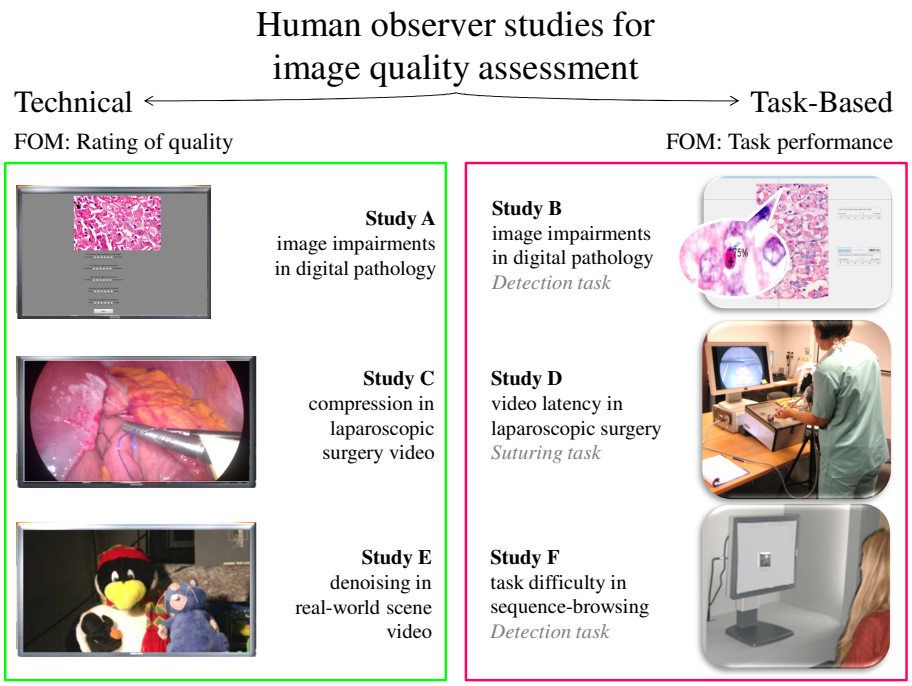


Figure 2.1: An overview of the psychophysical experiments for image/video quality assessment conducted in the course of this thesis. The author of this thesis acted as the principal and coordinating investigator in Study A (Section 2.4), Study B (Section 2.5), and Study F (Section 4.6); the others were co-principal roles. See also Table 2.1 for more information about the study parameters.

- dimensionality and presentation of image data
  - static 2D images (digital pathology slides);
  - moving 2D images (real-world scene videos, laparoscopic surgery videos);
  - 2D images viewed in a sequence-browsing mode (synthetic tomography-like image sequences).

In the next sections, we introduce in more detail the problem of IQA for digital pathology which served as the motivation for two of our observer studies: the study of TechIQ, hereafter referred to as Study A (see Section 2.4) and the study of TaskIQ, hereafter referred to as Study B (see Section 2.5). Separately in Chapter 4, we present the study which examined the role of image parameters in static versus sequence-browsing data presentation. Furthermore, an additional observer study from this dissertation is presented in Chapter 6 investigating the quality of appearance rather than the TechIQ or the TaskIQ.

Table 2.1: Some details about the study design for the psychophysical experiments from Figure 2.1. Further details concerning studies A and B are presented respectively in Section 2.4 and Section 2.5. The key resulting publications so far include: for Study A [Platiša et al., 2013b], for Study B [Platiša et al., 2013a], for Study C [Kumcu et al., 2014], for Study D [Kumcu et al., 2013], and for Study F [Platiša et al., 2012c, Kumcu et al., 2012b].

IQA approach	Study name	Image data	Reader expertise	Num readers	Num cases per reader	
					single-stimulus	*double-stimulus*
TechIQ	A	*2D digital pathology*	Diagnostic pathologists	6	72	0
			Veterinary students	7		
			IPI researchers	11		
	C	Laparoscopic video	IPI researchers	16	20	2 x 70
			Laparoscopic surgeons	5+		
	E	Real-world scene video	IPI researchers	18	20	2 x 70
			UGent students	16		
TaskIQ	B	*2D digital pathology*	Diagnostic pathologists	6	72	0
	D	Laparoscopic video	Laparoscopic surgeons	32	12	0
	E	*Computer generated sequences*	IPI researchers	22	370 (2D)	0
					370 (3D)	0

2.2.2 IQA for digital pathology slides

Presently, traditional microscopy is undergoing a major transformation driven by the development of automated whole slide imaging[1] [Weinstein et al., 2009] (WSI). The advent of new WSI systems creates the need for research on IQA methodology for these systems and development of adequate (application-specific) perceptually relevant IQ measures [Yagi and Gilbertson, 2005]. While the field of IQA for natural scene images and videos has advanced remarkably in the last years [Wang, 2011], the development of IQ measures suited specific image domains such as digitized artworks [Polatkan et al., 2009, Farnand et al., 2009] (studied in Chapter 6), is still at its infancy. In the area of medical imaging, some progress has been made in the field of advanced diagnostic imaging (MRI, CT, nuclear medicine imaging) [Reiner, 2013] and more so in the domain of digital mammography [Young et al., 2010]; related IQA techniques are surveyed in Chapter 3 and Chapter 4. Nevertheless, many areas of medical imaging, including digital pathology, still require research into the development of appropriate indicators for IQ.

Despite the booming popularity and the advance in new technologies for the WSI [Rojo et al., 2006], at this moment there is still no standardization regarding validation of digital pathology for diagnostic purposes [Lange, 2011, Pantanowitz et al., 2011]. It was only very recently that a "Guideline from the College of American Pathologists Pathology and Laboratory Quality Center" appeared [Pantanowitz et al., 2013]. According to [Henricks, 2012], wider adoption of WSI in pathology practice is anticipated to occur following further technical advancements (*e.g.* quality control of the process, time to prepare and scan a slide, scanning failure rate) as well as procedural advancements (*e.g.* clinical workflow, standardization). Also impacting the adoption of WSI is the current policy of the US Food and Drug Administration to consider WSI systems as class III (highest risk) medical devices [Yagi and Pantanowitz, 2012]. In line with the aims of this thesis, our focus is on the issues of measuring IQ for digital pathology images.

The logical initial steps to defining any measure are, firstly, to identify the key influencing factors and, secondly, to evaluate their relative importance for the observed output. In our case, the factors are common types of image manipulation and image impairment (hereafter jointly referred to as $manipulation$) in digital pathology slides; further details follow in Section 2.3.2. The output is twofold, depending on the IQA approach:

1. Study A: the perceived IQ attributes (P-IQ-attributes), and

[1] Whole slide imaging (WSI), commonly also termed "virtual" microscopy, refers to the process of digitization of glass slides, either the entire slides or the user selected area of it. Current WSI devices are capable of automatically producing high resolution digital images of high magnification (*e.g.* 40 times) within a relatively short time (on the order of minutes, depending on the slide and scanning parameters). The main advantages of digital over conventional light microscopy include: ease of access and sharing of images including remotely, reproducibility, and use of automated image analysis (computer-aided diagnosis (CAD) systems). Nevertheless, it is yet to be demonstrated if pathologists can be as effective with WSI as they are with the optical microscope [Redondo et al., 2012, Gallas et al., 2013]

2. Study B: the diagnostic performance of observers when interpreting the images,

both in the case where the images are viewed on a medical color liquid crystal display (LCD). Specifics of the considered manipulations, P-IQ-attributes, and the diagnostic task are described in the following sections.

In Study B, the observer is asked to perform a clinical diagnosis which requires specialized expertise and thus the study is restricted to the experts in the field. Conversely, Study A could possibly also be performed by non-expert observers (naive to diagnostic pathology). Commonly, histopathological data are used by only a very limited number of professional pathologists who are often not readily available and rather costly as participants in subjective IQ evaluations. Therefore, it would be of great benefit if pathologists could be replaced by naive or non-pathology expert observers while at the same time retaining the practical relevance of the study results. This is especially of interest for the larger-size studies [Liu et al., 2012]. The question of expertise becomes more relevant in light of studies of expertise in pathology [Krupinski et al., 2006, Mello-Thoms et al., 2011, Krupinski et al., 2013]. Evidence suggests that, alongside diagnostic performance, also visual search strategies significantly change as a function of level of experience, *i.e.*, as trainees become more familiar with the expected image content and which image details and characteristics are indicators of relevant information for rendering diagnostic decisions. Even in the domain of natural scene images, recent investigations suggest that visual attention (saliency distribution for an image) is influenced both by image content [Liu et al., 2013] and by the task of assessing the quality (rather than only viewing the images) [Gide and Karam, 2012]. It would not be unreasonable to assume that some or all of these aspects also influence the IQ judgments of humans, and that perhaps differently for different expertise profiles.

2.3 Test images: Digital pathology slides

The reference (undistorted) images in our experiments are real digital pathology data provided by Dr. Leen Van Brantegem from the Laboratory of Veterinary Pathology, Department of Pathology, Bacteriology and Avian Diseases, Faculty of Veterinary Medicine at Ghent University. To create the distorted images, we applied on the reference images a range of controlled computer manipulations, such as blur filtering or compression. The details are described next.

2.3.1 Reference images

All images were crops of digital pathology slides of 3 different animal tissue samples (Tiss1, Tiss2, Tiss3), each stained with haematoxylin and eosin (H&E) following the same procedure. The images were potentially showing pathological conditions characterized by inclusion bodies (hereafter *lesions*). The images were all 1200×750 pixels in size. In the main study, a total of 12 non-manipulated images (hereafter

denoted "M-NONE" and commonly also referred to as *reference* images), 4 of each Tiss1, Tiss2, and Tiss3 were used; these are shown in Figure 2.2. Out of these 12 reference images, 5 images were normal (lesion-absent) cases and 7 images were abnormal (lesion-present) cases which contained one or two lesions; for further explanation see Section 2.5.3.

2.3.2 Image manipulations

We are interested in studying the following common factors of image acquisition, management, and displaying within the WSI systems: *blurring* (possibly caused by thick or folded tissue, incorrect focus, vibrations during scanning), *color* and *gamma* parameters (typically controlled by the parameters of the display system), *noise* (possibly lower for live tissue and higher for dead tissue samples; increasing when the microscope approaches the resolution limit), and image *compression* (necessary for storage and especially transmission of the very large sizes of digital pathology images).

For the purpose of studying these effects, the reference images were artificially altered by: adding Gaussian blur ($\sigma_b = 3$), unsharp masking, decreasing/increasing gamma (approximately -5%/+5%), decreasing/increasing color saturation (approximately -5%/+5%), adding white Gaussian noise ($\sigma_n = 10$), and JPG compression (libjpeg[2] quality 50). The manipulations were applied on each reference image and always one at a time. The degree of each manipulation was selected in the grayscale (luminance) domain such that all degraded images had approximately the same amount of perceived difference relative to the corresponding reference image – subtle yet noticeable. The degree of perceptual difference was measured using the High dynamic range visible difference predictor (HDR-VDP) [Mantiuk et al., 2005].

In a pre-study experiment, the following five image manipulations were selected to have the most prominent effect on the P-IQ-attributes:

1. added Gaussian blur ($\sigma_b = 3$),
2. decreased gamma (approximately -5%),
3. decreased color saturation (approximately -5%),
4. added high-frequency white Gaussian noise ($\sigma_n = 10$), and
5. JPG compression (libjpeg quality 50).

More details about the pre-study experiment design are available in Section 2.4.2. For conciseness, however, the full details of the pre-study results are not reported here.

Further in the text, the aforementioned five categories of manipulated images are denoted by an "M-" prefix, for example, we write "M-Blur" to denote an image which was manipulated by adding Gaussian blur of $\sigma_b = 3$. Also, we write "M-NONE" to

[2] `http://libjpeg.sourceforge.net/`

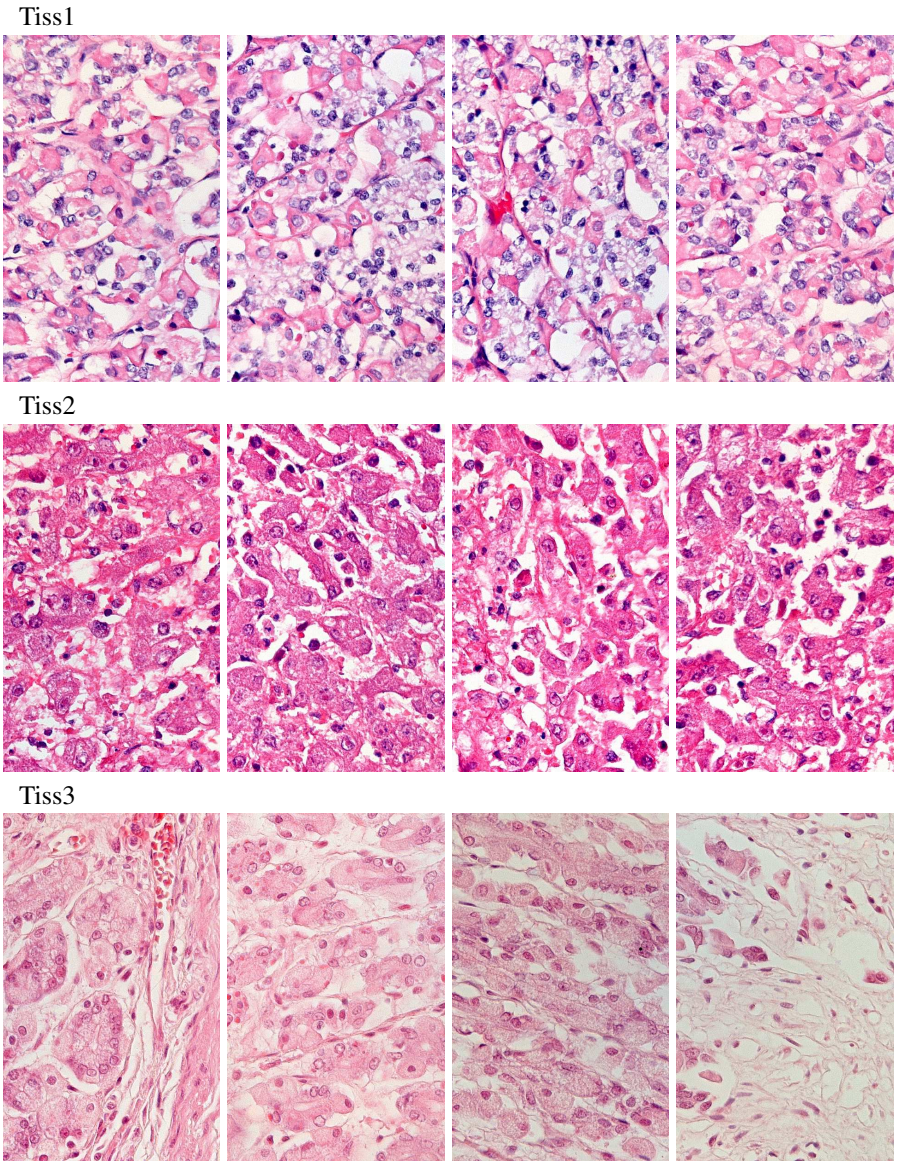


Figure 2.2: The 12 reference (non-manipulated/unimpaired, M-NONE) images: 4 different crops are taken from each of the 3 considered classes of pathological slides: Tiss1 - gastric fundic glands of a dog (top), Tiss2 - liver of a foal (middle), and Tiss3 - gastric fundic glands of a dog (bottom). All 12 M-NONE images, as well as all their $12 \times 5 = 60$ artificially transformed variants (M-Blur, M-Gamma, M-ColSat, M-Noise, M-JPG) are 1200×750 pixels in size.

denote a reference image and "M-Any" to refer to any test image but the M-NONE (when the exact type of manipulation is not of interest). Thus, the main study dataset included a total of 72 images: 12 M-NONE images, 12 M-Blur images, 12 M-Gamma images, 12 M-ColSat images, 12 M-Noise images, and 12 M-JPG images. The same dataset was used in both Study A (Section 2.4) and Study B (Section 2.5).

2.4 Study A. Perceived quality of the images

We now present the details of the first human observer study which followed the TechIQ approach to IQA. In the following, we describe: (1) the experimental goal, (2) the design of the experiments (including the numbers of test images and participating observers, the training process, the conditions of image viewing, and the experimental questions), (3) the methods of data analysis, and (4) the results of that analysis (both qualitative and quantitative).

2.4.1 Experimental goal

In this study, we address two main questions: (1) What is the relationship between image manipulations and P-IQ-attributes? (2) How is that relationship influenced by the expertise profile of the observers? In particular, we consider three groups of observers: experts in diagnostic veterinary pathology (pathology experts, PExperts), students of veterinary medicine (pathology students, PStudents), and researchers in digital image processing (imaging experts, IExperts). In addition to revealing the associations between different expertise groups, the outcomes of this study could be used to suggest directions for optimizing the related imaging systems (e.g. training versus clinical systems) for specific users (e.g. trainees versus practicing clinicians). For example, we could enhance a particular quality attribute which is more significant for a given category of users, while investing less in a less influential feature.

2.4.2 Study design

As previously mentioned, the main study was preceded by a pre-study which narrowed the selection of image manipulations to those with dominant perceptual effects. The two studies differ in terms of observers (human subjects) and test images, and they are the same in terms of the experimental task and environment.

A total of 24 observers participated in the main study: 6 PExperts, 7 PStudents, and 11 IExperts. Details about the gender, age and experience distributions for each expertise group are summarised in Table 2.2. All observers were screened for color vision deficiencies using the Farnsworth Panel D15 test [Farnsworth, 1947] and they were all found not color blind.

Each observer evaluated a total of 72 images: 12 reference images and their 12×5 manipulated variants. All image evaluations were single-stimulus (SS) trials as de-

Table 2.2: Distribution of the observers according to gender, age, and average experience in diagnostic pathology, by expertise group

	Num All	Num Male	Num Female	Min / Max Age	Median Age	Mean Experience
PExperts	6	1	5	25 / 40	29.5	6.2
PStudents	7	0	7	21 / 28	22.0	2.0
IExperts	11	8	3	24 / 36	29.0	5.5

scribed in ITU-R Recommendation BT.500-13 [ITU-R, 2012] in which a single randomly chosen image was displayed at a time. Note that no distinction was made in the presentation of a reference versus a manipulated image, *i.e.*, there was never any explicit information given to the observer about the presence nor about the type of manipulation for a displayed image (hidden reference) [Redi et al., 2010]. For each image, the following five P-IQ-attribute ratings were collected: perceived overall IQ (P-IQ), perceived blur disturbance (P-Blur), perceived quality of contrast (P-Contrast), perceived noise disturbance (P-Noise), and perceived quality of color saturation (P-ColSat). Each attribute was rated using a 6-point absolute category rating scale [ITU-R, 2012] ranging from 0 to 5. The better (less disturbing) the perceived attribute, the higher the score. An example screen capture of the proprietary web-based interface used for displaying images and collecting observer responses is shown in Figure 2.3.

The images were displayed on a 3MP medical color LCD display (MDCC-3120-DL, Barco N.V., Kortrijk, Belgium) with the color management set to fidelity. No image adjustment (zoom, window level) was allowed. The observers were seated at 50 cm from the display and were allowed to lean back and forth. The experiments were conducted in a controlled viewing environment to ensure consistent experimental conditions: low surface reflectance and approximately constant ambient light. There was no time limitation.

Each observer evaluated images in a session which comprised a *training* and a *rating* phase. The training phase began with a combined written and verbal description of the study goals and its design, continued with a small hands-on tutorial about the attributes of IQ (for PExperts and PStudents), and ended with 10 image trials (not considered in the data analysis) which were aimed at familiarizing the observer with the images and the range of IQ in the experimental data as well as with the graphical user interface and the rating scales (see Figure 2.3). We note that (due to practical constraints of the project) the number of the training images was rather small which raises the risk of larger variability of the human ratings. We discuss this further in Section 2.4.4.

After the training part, which took on average between 10 (IExperts) and 20 minutes (PExperts and PStudents), the rating part began. The results presented in Section 2.4.4 are generated using the attribute ratings collected in the rating part of the main study.

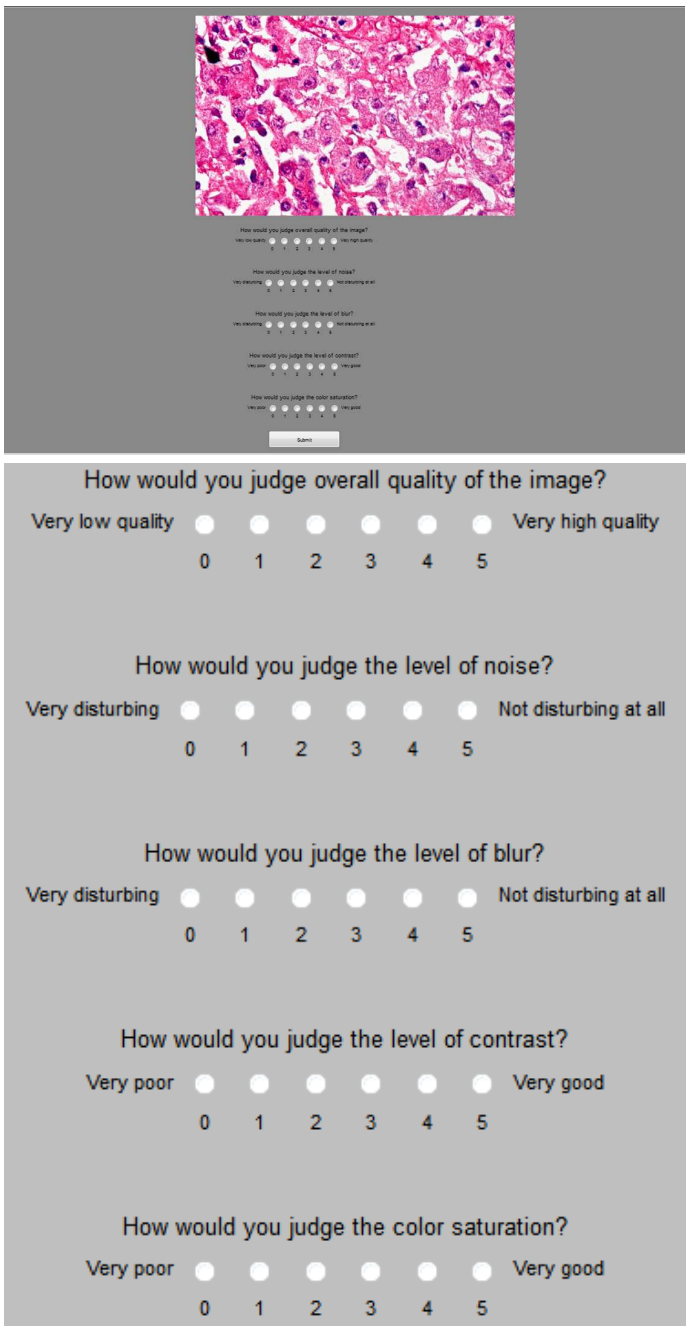


Figure 2.3: Graphical user interface for Study A. (Top) Screen capture of a trial from the experiment. (Bottom) Zoom in of an area on the screen showing the test questions and their associated rating scales.

Additionally, 4 observers participated in a pre-study experiment but not in the main study: 1 PExpert, 2 IExperts and 1 PStudent. The pre-study was conducted for the initial 9 types of image manipulation involving a total of 30 images: 3 reference images and their 3×9 manipulated variants (see Section 2.3 for details on the types of image manipulation).

2.4.3 Technical FOM

As previously described, the observers rated IQ attributes in single-stimulus trials using a 6-point rating scale from 0 (very low IQ) to 5 (very high IQ). The data analysis is performed in line with the ITU-R Recommendation BT.500-13 [ITU-R, 2012].[3]

First, we performed inconsistency testing per subject by comparing the collected data of individual observers to those of other observers from the same expertise group. None of the observers was determined inconsistent.

Next, we examined the descriptive statistics of our collected data to ascertained whether the distribution of observer ratings (scores) is normal or not. Typically, assuming a normal distribution of scores for each test condition, single-stimulus data is analyzed using the mean opinion score (MOS) obtained by rejecting outliers and computing the mean of all observer ratings over a stimulus [Wang et al., 2004b, ITU-R, 2012]. However, our analysis suggested non-Gaussian distribution. Therefore, we have chosen to use the $median$ opinion score (MdnOS) and the [25%, 75%] interquartile range (IQR) to evaluate observer ratings. To test for differences between MdnOSs, we perform the Kruskal-Wallis[4] non-parametric one-way analysis of variance (ANOVA) and post-hoc pair-wise comparisons at a significance level $\alpha = 0.05$.

2.4.4 Results and discussion

Figure 2.4 shows the boxplots of the quality ratings gathered in our main study. On each box, the dot represents the MdnOS, the length of the box represents the IQR, the whiskers extend to 1.5 IQR, and "+" marks denote "outliers" (measured points outside of the whisker range). We use color to distinguish different observer groups: red for PExperts, blue for PStudents, and green for IExperts. The rows in the figure correspond to the five P-IQ-attributes (P-IQ, P-Blur, P-Contrast, P-Noise, P-ColSat) and columns represent the three tissue types (Tiss1, Tiss2, Tiss3). The six different

[3] An alternative approach to data analysis (not explored here) could be based on the framework of multidimensional scaling (MDS) as in [Vansteenkiste et al., 2006]. The MDS approach employs the multidimensional geometric model to describe the relationships among different attributes of IQ as well as between the attributes of IQ and the overall IQ. The dimensionality of the MDS model is determined by the number of independently varying attributes. More information on MDS models of IQ can be found in [Ahumada and Null, 1993, Martens, 2002].

[4] We note that the data in our experiments do not always comply to the condition of data independency which is required for the Kruskal-Wallis analysis, *e.g.*, the impaired images in our experiments are manipulated variants of the reference images and thus M-Any images are related to M-NONE images. Nevertheless, for simplicity reasons and given the preliminary character of the study, we will consider the images of different M-groups (approximately) independent.

types of image manipulations (M-NONE, M-Blur, M-Gamma, M-ColSat, M-Noise, M-JPG) are indicated on the x-axis of each graph.

Before analyzing the effects of image manipulations, we will first briefly look at the perceptual scores of the reference (M-NONE) images across the three classes of content (Tiss1, Tiss2, Tiss3). Next, we turn to the image manipulations and associated effects. In particular, we examine two types of effects: "effect of image manipulation" and "effect of observer expertise". With the effect of manipulation, we refer to the change in the P-IQ-attribute for an M-Any relative to the M-NONE image for each observer group and each content class independently. The effect of expertise, on the other hand, examines the difference in ratings between observer groups (PExperts-PStudents, PExperts-IExperts, PStudents-IExperts) at a given P-IQ-attribute (for a given content class and a given manipulation).

2.4.4.1 Qualitative analysis

We start from the boxplots shown in the first row of Figure 2.4 which represent the P-IQ ratings. Simply by visual comparison, we notice that PExperts made a rather clear ranking of the 3 contents: Tiss1 images were rated the highest quality (MdnOS$\approx$4), next were Tiss3 (MdnOS$\approx$3) and lowest were Tiss2 (MdnOS$\approx$2). The ratings by IExperts were relatively similar for all 3 contents (MdnOS$\approx$3) though with some variations in the corresponding IQRs. Lastly, the results of the PStudent group lay somewhere in between those of the PExperts and the IExperts: they could discern the differences in IQ of the 3 content classes better than the IExperts but not as clearly as the PExperts. These differences in judgments of the overall IQ of the 3 classes of reference images, especially in the case of PExperts versus IExperts, possibly suggest that PExperts and IExperts judged IQ with different minimum expectations. Namely, if we look at the boxplots of the other 4 considered P-IQ-attributes (still focusing on M-NONE images only), we notice that for Tiss1, for example, PExperts found all IQ attributes of a relatively high quality (MdnOS$\geq$4) while IExperts perceived noise to be rather disturbing (MdnOS$\approx$3). Even more PExperts versus IExperts disagreements over specific P-IQ-attributes can be observed with Tiss2: the two expert groups seem to disagree most about P-Blur but also about P-Contrast and P-Noise, and more importantly, they seem to have different criteria for judging the impact of different P-IQ-attributes on the P-IQ. More details about the significance of the observed differences between different observer groups are given later in this Section when discussing Table 2.4.

2.4.4.2 Effects of image manipulation

Table 2.3 depicts the results of the analysis for effect of manipulation. Columns in the table denote the type of image manipulation, grouped by three classes of content (Tiss1, Tiss2, Tiss3). Observer groups are represented by rows and grouped by the judged P-IQ-attribute. Cross marks are used to denote manipulations which resulted in

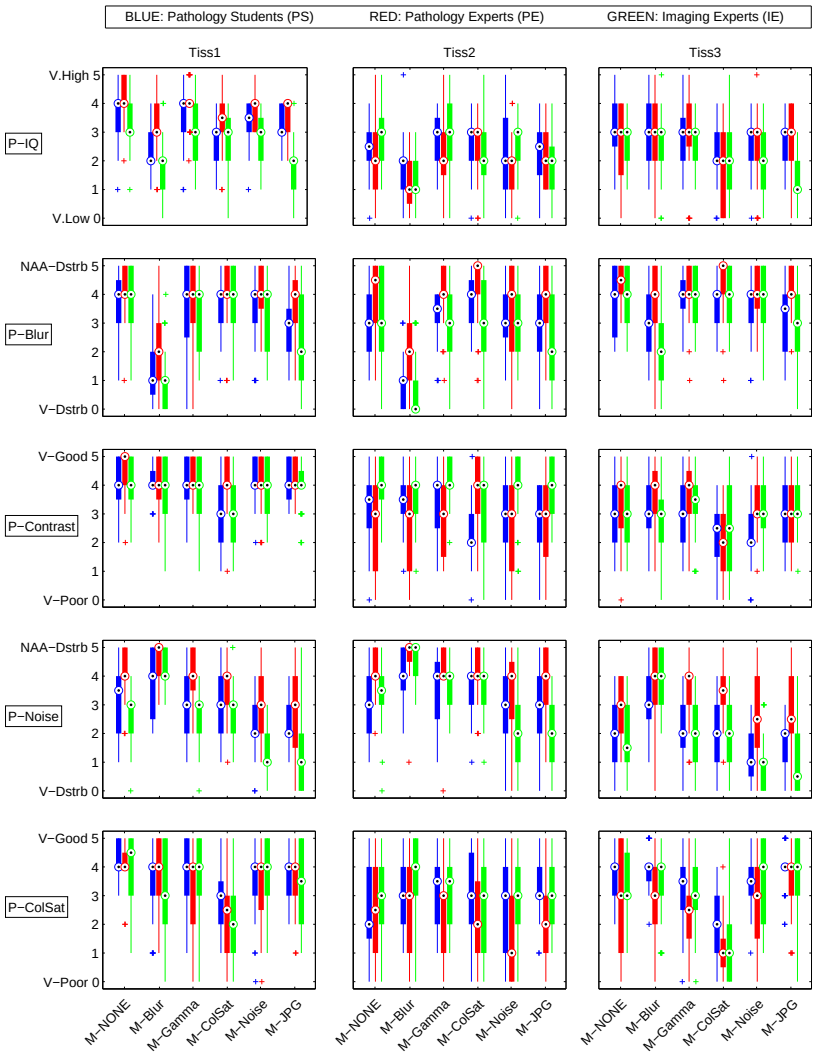


Figure 2.4: Boxplots of P-IQ-attribute ratings. Each box indicates the median (MdnOS), the interquartile range (IQR), the 1.5 IQR interval (whiskers), and the "outliers" (measured points outside of the whisker range denoted by "+" marks). Each boxplot corresponds to one of the 3 tissue classes (columns, from left to right): Tiss1, Tiss2, Tiss3 and one of the 5 judged P-IQ-attributes (rows, from top to bottom): P-IQ, P-Blur, P-Contrast, P-Noise, and P-ColSat (see text for details). For each boxplot, the scores were grouped by 3 observer groups (PExperts, PStudents, IExperts) and by 6 types of image manipulation (M-NONE, M-Blur, M-Gamma, M-ColSat, M-Noise, M-JPG). All P-IQ-attributes were rated using a 6-point grading scale from 0 to 5 (0 - very poor/very disturbing, 5 - very good/not disturbing at all).

statistically significant P-IQ-attribute ratings compared to the corresponding reference (M-NONE) images. Based on Table 2.3, we make the following observations.

1. Irrespective of the expertise, most observers judged that manipulations did not affect P-IQ significantly, especially not for Tiss2 and Tiss3 contents. Manipulations which did have some effect (though not for all contents) are blurring and decrease in color saturation. Interestingly, IExperts almost always rated P-IQ of JPG compressed images significantly lower compared to the reference. On the other hand, PExperts and PStudents were little sensitive to JPG compression. We return to this suggested PExperts insensitivity to JPG compression in Section 2.6.

2. All expertise groups perceived blur as disturbing. The only exception were PExperts and PStudents in the case of Tiss3. One possible explanation for them being less disturbed by blur in Tiss3 could be in the fact that the edges in Tiss3 images were overall less prominent (less contrast) than in Tiss1 and Tiss2 and so blurring of the edges was less obvious, especially to a non-imaging expert with limited training (see Section 2.4.2). This reasoning about the lack of expertise is also in line with the P-Blur plots in Figure 2.4 which show a large spread of the P-Blur ratings not only for PExperts and PStudents but for IExperts as well.

3. Except for the decrease in color saturation, and then only sporadically, none of the manipulations had a major effect on P-Contrast.

4. Perhaps somewhat surprisingly, P-Noise in the images with added noise (M-Noise) was found significantly more disturbing than that of the reference images only by IExperts, but not by PExperts, nor by PStudents. This suggests that PExperts could be little sensitive to noise, possibly because noise does not interfere much with their clinical use of the images. That is, it could be that other image attributes (such as sharpness, color, contrast) are more important for the correct interpretation of histopathological images and hence PExperts find noise less disturbing than "unbiased" observers such as IExperts.

5. Reduction of color saturation did have an effect on P-ColSat, for all PExperts, PStudents and IExperts, and for both Tiss1 and Tiss3. No significant difference was captured for Tiss2, which is not surprising since the reference images of this content class were already highly saturated in color, and (from Figure 2.4) they were rated more disturbing for P-ColSat than the Tiss1 and Tiss3 reference images.

2.4.4.3 Effects of observer group

We now turn to examine the effects of observer expertise in more detail. The results of the related data analysis are summarized in Table 2.4. Again, columns in the table

Table 2.3: Effect of image manipulation: For a given content class (Tiss1, Tiss2, Tiss3), are the ratings for M-Any and M-NONE significantly different? PE, PS, and IE denote the observer groups, respectively: PExperts, PStudents, and IExperts. Cross marks denote significant effects (see text for details).

		Tiss1						Tiss2						Tiss3					
		M-NONE	M-Blur	M-Gamma	M-ColSat	M-Noise	M-JPG	M-NONE	M-Blur	M-Gamma	M-ColSat	M-Noise	M-JPG	M-NONE	M-Blur	M-Gamma	M-ColSat	M-Noise	M-JPG
P-IQ	PE	-	x		x			-						-					
	PS	-	x		x			-						-			x		
	IE	-	x				x	-	x				x	-					x
P-Blur	PE	-	x					-	x					-					
	PS	-	x					-	x					-					
	IE	-	x				x	-	x					-	x				
P-Contrast	PE	-						-						-			x		
	PS	-			x			-						-					
	IE	-			x			-						-					
P-Noise	PE	-						-						-					
	PS	-						-						-	x				
	IE	-	x			x	x	-	x			x	x	-	x			x	
P-ColSat	PE	-			x			-						-			x		
	PS	-			x			-						-			x		
	IE	-	x		x			-						-			x		

denote the type of image manipulation, grouped by 3 classes of content (Tiss1, Tiss2, Tiss3). Rows represent pairs of observer groups which are being compared, specifically: PExperts-IE, PExperts-IExperts and IExperts-PStudents, and they are grouped by the judged P-IQ-attribute. Cross marks are used to denote manipulations which resulted in statistically significant differences between the P-IQ-attribute ratings of the two paired observer groups. For convenience, we will refer to the set of study parameters corresponding to each cell in Table 2.4 as study setup (type of image manipulation, perceived attribute).

Overall, we notice that PExperts and IExperts differed for their ratings most often, PExperts and PStudents differed on notably fewer occasions, while IExperts and PStudents differed in only a few scenarios. The position of PStudents in this arrangement (being inconsistent with either PExperts or IExperts) could perhaps be interpreted as a more conservative approach of the PStudent group, or it might simply reflect the transition between a non-pathologist to an expert pathologist. Overall, the differences between the groups could be caused by multiple reasons, including the following: (1) the limited training and possible confusion about the different types of artifacts

(especially of interest for non-IExperts), (2) the nature of the experimental question (how disturbing is the impairment?) – it is not unexpected that imaging experts are more disturbed by image impairments than other observer profiles, and possibly also (3) the difference in the total number of images viewed (PExperts already viewed the images in Study B, see Section 2.5). For the following analysis, we will focus on PExperts-IExperts comparisons.

The most striking difference between the two expert groups are observed for P-Noise ratings. From Figure 2.4, row 4, we notice that P-Noise was usually rated more disturbing by IExperts than by PExperts. This is in agreement with the previous discussion of Table 2.3 which argued that PExperts might be less sensitive to noise than IExperts.

Next, we find that P-IQ ratings of PExperts and IExperts were different for Tiss1; IExperts always rated the quality lower than PExperts. This is in line with the discussion about P-IQ for M-NONE images earlier in this Section: IExperts seem much more disturbed by noise than PExperts, even for the reference images of this content class. The PExperts on the other hand, appear nearly insensitive to the noise artifacts.

Finally, PExperts tend to be notably more sensitive than IExperts to the changes in color saturation and gamma. For example, PExperts ratings for P-ColSat were consistently lower than those of IExperts when gamma was decreased, and often also when color saturation was decreased. Thus, adequate settings and perhaps even options for adjusting the display parameters of color and gamma appear to be of great importance for PExperts.

2.5 Study B. Diagnostic performance on the images

In this section we present the second human observer study focusing on the TaskIQ - based image assessment. The same as for the previous study, we consider the following: (1) experimental goal, (2) design of the experiments, (3) methods of data analysis, and (4) study results.

2.5.1 Experimental goal

The paradigm of task-based IQA holds that the images used for a specific task should be evaluated based on how well they fit the purpose. For example, medical images are meant to serve as a diagnostic or a medical intervention tool. Thus, the best way to assess the IQ of medical images/imaging systems is by measuring the level of performance of medical doctors (the intended end-users) while they are using the given images to accomplish the task of interest (*e.g.* interpreting mammogram images in order to distinguish benign from malignant tissue).

Our goal in this study is the same as in Study A (Section 2.4): to find out whether there is an effect of any of the considered image manipulations on the quality of veterinary digital pathology images. However, in contrast to Study A in which we were

Table 2.4: Effect of observer expertise: Does observer expertise have an effect on the P-IQ-attribute rating for a given M-Any? PE, PS, and IE denote the observer groups, respectively: PExperts, PStudents, and IExperts. Cross marks denote statistically significant effects (see Section 2.4.3 for details).

		Tiss1						Tiss2						Tiss3					
		M-NONE	M-Blur	M-Gamma	M-ColSat	M-Noise	M-JPG	M-NONE	M-Blur	M-Gamma	M-ColSat	M-Noise	M-JPG	M-NONE	M-Blur	M-Gamma	M-ColSat	M-Noise	M-JPG
P-IQ	PE - PS																		
	PE - IE	x	x	x	x	x	x	x				x							x
	IE - PS		x			x	x												
P-Blur	PE - PS							x		x			x	x					
	PE - IE		x				x	x	x	x	x		x	x	x				x
	IE - PS																		
P-Contrast	PE - PS				x						x					x		x	
	PE - IE				x			x		x		x	x		x				
	IE - PS							x		x	x	x	x					x	
P-Noise	PE - PS	x	x	x	x			x	x	x			x	x		x	x	x	x
	PE - IE	x		x	x	x	x	x		x		x	x	x		x	x	x	x
	IE - PS					x	x		x										
P-ColSat	PE - PS											x			x		x		
	PE - IE			x		x				x		x	x			x			
	IE - PS																		

interested in TechIQ rating, here we aim to quantify the TaskIQ, *i.e.*, the effects of the manipulations on the utility of images. In the other study, the observers were asked exclusively to rate the quality of images (either the overall IQ or the attributes of IQ). In this study, we ask the observers to perform a certain task using the images under test. If we would find that the level of performance for some manipulations is lower than for the unimpaired (reference) images, that would indicate that the practical utility of those impaired images is less compared to the unimpaired images. Under the TaskIQ approach, the lower practical utility corresponds to the lower quality of those impaired images.

2.5.2 Study design

The study was designed in consultation with an expert diagnostic veterinary pathologist. The test images were exactly the same as in the study for TechIQ assessment from Section 2.4 but the observer's task was different and it comprised the following two steps:

1. Judge the overall quality of the image using a continuous scale[5] from 0 (low quality) to 100% (high quality).

2. Mark and rate all suspected color abnormalities (technical term: inclusion bodies; in this text referred to as lesions), knowing that any number of them is possible, including zero lesions (also called a lesion-free or a normal image).

The study followed the fully-crossed multi-reader multi-case (MRMC) paradigm, *i.e.*, every observer (reader) interpreted every image (case). Considering the fact that the test images may contain an arbitrary number of abnormalities, we opt for the so-called free-response receiver operating characteristic (FROC) design of an MRMC study in which the task for the reader is to detect and locate each suspected abnormality.[6] In particular, the observer was asked to mark every suspected location which they consider worthy of mention [Metz, 2006] and rate their confidence of abnormality using a continuous scale from 0 (low confidence) to 100% (high confidence).

Given the observer's task, which requires highly specialized knowledge and skills in the field of diagnostic veterinary pathology, this study was conducted only by the PExpert group of observers, as they had the necessary expertise. The observers were exactly the same 6 PS from Study A. They all first completed the experiment from Study B and later the experiment from Study A (never on the same day but usually a few days later). This is important to note because, unlike in Study A where the PExperts received some "training" about the attributes of IQ (see Section 2.4 for details), in Study B there were no instructions provided in that sense; rather the observers were asked to judge the IQ according to their own personal criteria.

Specifically, the task instructions for Study B (provided in written form and explained verbally) covered the following aspects:

- motivation for the study (to assess the effects of color image quality on diagnostic performance),

- brief description of the image data (types of imaged tissues, staining protocol, number of images, none of the presented images are exactly the same),

- description of the task (as already described),

[5] The continuous scale was suggested by our consulting expert pathologist as preferred (more comfortable) over the discrete one.

[6] Three state-of-the-art paradigms are commonly used to assess and compare the diagnostic performance in MRMC studies for joint location and detection of lesions: the location ROC (LROC) [Starr et al., 1975], the free-response ROC (FROC) [Bunch et al., 1977, Chakraborty, 1989, Chakraborty and Berbaum, 2004], and the region of interest (ROI) approach [Obuchowski et al., 2000]. As suggested by their names, all these methods rely on the basic principles of conventional ROC analysis [Metz, 2006] which we discuss in Chapter 3 of this dissertation. The ROI approach a priori defines image regions (segments of the actual images, each with one or more pathologies) and treats them (rather than images as a whole) as the basic elements of ROC analysis; such treatment of the data is not directly suited for our study. In contrast, the LROC and FROC approach both treat the image as a whole and they differ in the number of allowed lesions per image; the LROC restricts the number of possible lesions to only one per image while the FROC allows any number of lesions per image. A concise and critical overview of the three classes of methods can be found in [Zhou et al., 2011].

- explanation of the graphical user interface (mark new lesion, cancel the last lesion, assign a confidence rating, proceed to the next image),
- expected accuracy in annotation (the center of a suspicious region needs to be marked with reasonable accuracy, otherwise it is considered an error),
- basic principle of the performance analysis (the credit is given to the correctly marked lesions with high ratings while highly rated false positives are penalized), and
- privacy statement (the results will only be used in de-identified form and will not be revealed to others).

Before interpreting the test images, the observers have interpreted a separate set of 5 images for training purposes (those results are not considered in the data analysis); no feedback was provided. The same as in Study A, the number of training images was rather limited (for the same reasons of time and resource constraints on the project and the preliminary character of the studies). Nevertheless, the observers were all very familiar with the task and we expected to see no major effects of the potential lack of training (such as large variability in the observer scores). Moreover, in the questionnaire which the observers filled-in after completing the experiments, all PExperts indicated that the training was "sufficient". We discuss this further in Section 2.5.4.

The 72 test images described in Section 2.3 were presented as follows: the $12 \times 5 = 60$ impaired images were shown first (randomly ordered) followed by the unimpaired 12 images (randomly ordered). Showing the unimpaired images last aimed at excluding the bias of "learning" from those. The viewing conditions were the same as in Section 2.4. The duration of the experimental session was about two hours, including all the steps - the instructions, the training, the experiment and optional breaks, and the short questionnaire about professional experience and realization of the experiment.

2.5.3 Diagnostic FOM

Consistent with the FROC study design, the collected human data consists of an arbitrary number of mark-rating pairs per image. Under the FROC paradigm, a marked (suspected) location is classified as a correct lesion localization (true positive, TP) if the mark falls within an *acceptance region* of the actual lesion; otherwise, it is a wrong lesion localization (false positive, FP). Note that the classification category, TP or FP, refers to a single marked location within an image and not to the image as a whole (as it is the case in ROC studies). Thus, in general, there can be an arbitrary number of TPs and FPs per image.[7] In our study, the acceptance region was defined as a manually delineated rectangular area determined by the largest width and height of

[7] Also conversely to the ROC studies, the concept of a negative (true negative, TN, or false negative, FN) is undefined and unmeasurable. [Chakraborty, 2010]

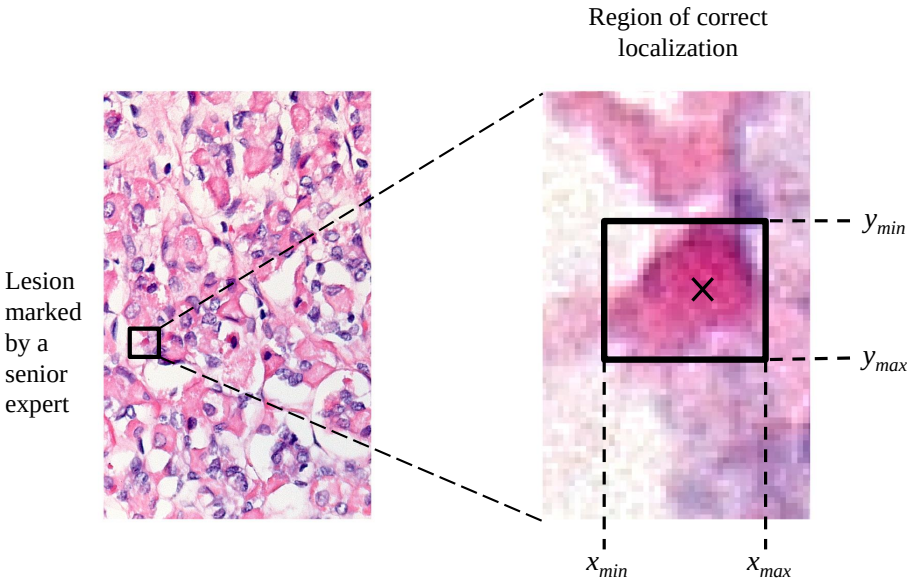


Figure 2.5: Location-level mark classification. A mark is accepted as "true-positive" if it belongs to the acceptance region around the actual lesion; otherwise, it is classified as "false-positive". The acceptance region is a manually delineated rectangular area determined by the largest width ($x_{max} - x_{min}$) and height ($y_{max} - y_{min}$) of the actual lesion in the reference image. The actual lesions are the lesions marked by the senior expert with confidence rating above 60 percent.

the actual lesion in the reference image. Here, the actual lesions were determined by an experienced senior expert in diagnostic veterinary pathology who only annotated the reference images but did not take part in the actual experiment; for illustration see Figure 2.5. We recall from Section 2.3.1 that 5 out of 12 reference images were normal (lesion-absent) cases while the remaining 7 were abnormal (lesion-present) cases, each containing one or two lesions.

After all marked locations are classified into TP or FP, the alternative FROC (AFROC) curve is formed by plotting the proportion of TP marks over all actual lesions (true positive fraction) on the y-axis and the fraction of images with 1 or more FPs on the x-axis, as the threshold for reporting a suspicious region is varied.[8] Quantitatively, the level of performance in an FROC task can be described as the area under the corresponding AFROC curve. For this purpose we use the jackknife AFROC (JAFROC) method proposed by [Chakraborty and Berbaum, 2004] and

[8]The FROC curve differs from the AFROC curve in that it plots on the x axis the average number of FPs per image. Consequently, the horizontal axis of an FROC curve is not normalized to a maximum value of 1. This creates difficulty in the cases where we want to compare, for example, diagnostic performance for two image sets with different ranges of FPs per image (*e.g.* images from two different imaging systems). To its advantage, the AFROC curve does not suffer from this problem.

refined in [Chakraborty, 2006]. The JAFROC FOM is the trapezoidal area under the alternative FROC (AFROC) curve, and the jackknife technique refers to the analysis of variance. The same as with the conventional ROC analysis, larger area under the AFROC curve means higher diagnostic accuracy of the system under analysis.

The data is analyzed using the freely available JAFROC software.[9] The JAFROC suggests using at least 50 image samples in order for the findings to be generalizable to the population of images (random-case analysis) and more than 3 observers in order to generalize to the population of readers (random-reader analysis). Otherwise, the analysis is valid only for the specific images/observers used in the study (fixed-case/fixed-observer analysis). Thus, given the numbers in our study (6 observers and 12 images per manipulation), we report the results of random-reader fixed-case analysis. Note that, unlike in Study A (Section 2.4) where different tissue types (Tiss1, Tiss2, Tiss3) were treated separately, the analysis in Study B is unaware of the type of the imaged tissue, that is, the images are grouped only based on the type of manipulation, regardless of the tissue type; this is due to the limited number of images.

2.5.4 Results and discussion

Firstly, we discuss diagnostic performance of the images (TaskIQ). Secondly, we analyze human responses to the experimental question of overall IQ (TechIQ).

2.5.4.1 TaskIQ. JAFROC analysis

The results of the JAFROC analysis are summarized in Table 2.5 and graphically represented in Figure 2.6. Overall, the null hypothesis that the 6 categories of image manipulations are equal for their performance in the considered diagnostic task (joint detection and localization of the lesions) is rejected at a 5% significance level ($F(5, 25) = 2.81, p = 0.0379$).

Specifically, statistically significant differences in the level of task performance are found in 3 out of 15 possible comparisons of image manipulations (see the shaded cells in the table): between the reference and JPG compressed (M-NONE and M-JPG), between images with decreased gamma and those with decreased color saturation (M-Gamma and M-ColSat), and between desaturated and JPG compressed images (M-ColSat and M-JPG). Note that the latter two comparisons involve color manipulations (M-ColSat). As described in Section 2.3, the levels of manipulations were selected using the non-color-aware HDR-VDP measure. Therefore, it is possible that the magnitude of perceptual difference between M-ColSat images and their reference images is larger (or smaller) than for the rest of manipulated images in the study. Because of this, we take with caution the results which refer to the M-ColSat images.

We focus here on differences in performance level relative to the reference image data, M-NONE. Such a difference is observed only for the JPG compressed images:

[9] Dev Chakraborty's FROC web site, `http://www.devchakraborty.com`

the accuracy in detection of the inclusion bodies is lower on compressed then on unimpaired images. This suggests that the JPG compression could cause degradation in the diagnostic performance and therefore may be not acceptable for clinical digital pathology; that is, in particular, for the diagnosis of inclusion bodies under the H&E staining. Clearly, this indication shall be verified with a more extensive study (more observers and especially more images) before any final conclusions are made. Nevertheless, our results make a compelling argument for further investigation in this direction.

Previously, [Marcelo et al., 2000] found no statistically significant difference between the diagnostic accuracy of non-compressed and that of JPG compressed images in telepathology. The same was concluded by [Seidenari et al., 2004], although they noted the intra-observer reproducibility in the diagnostic judgment to be lower for compressed images. In the domain of image analysis, [Nicolosi et al., 2012] concluded that JPG compression does not seem to significantly compromise the accuracy of angiogenesis quantification in the ovarian epithelial tumors. In contrast, [López et al., 2008] studied the effects of image compression on automatic quantification of immunohistochemical nuclear markers and found it to be dependent on the image content (number of cells per field, number/size of clusters) - the effect was small for low-complexity images (≥ 100 cells per field, without clusters or with small-area clusters) and substantial for high-complexity images ($< 35 - 50$ cells/field). Overall, it is important to note that these reports largely differ in their content of images, characteristics of lesions, diagnostic tasks under study (detection or quantification of lesions) as well as in the range of compression ratio/quality. That considered, it is perfectly legitimate to come to different conclusions for different image/task setups. In fact, any generalizations would be in conflict with the primary argument of the TaskIQ approach that the quality should be judged for a specific image set and a corresponding diagnostic task.

Concerning future research, it is necessary to also refer to JPEG2000, especially at higher compression ratios[10] [Parwani et al., 2011]. Similar to the studies of JPG compression, current literature reports about the effects of JPEG2000 [Cavaro-Ménard et al., 2013, Krupinski et al., 2012], albeit limited, suggest the need for future investigations to be directed at specific applications (anatomy, diagnostic task, compressions ratio of interest).

[10] Supplement 145 of the DICOM standards [DICOM, 2009] states the following concerning image data compression: "Because of their large size, WSI data are often compressed. Depending on the application, lossless or lossy compression techniques may be used. Lossless compression typically yields a 3X-5X reduction in size. The most frequently used lossy compression techniques are JPEG and JPEG2000. For most applications, pathologists have found that there is no loss of diagnostic information when JPEG (at 15X-20X reduction) or JPEG2000 (at 30X-50X reduction) compression is used. Lossy compression is therefore often used in present-day WSI applications. JPEG2000 yields higher compression and fewer image artifacts than JPEG; however, JPEG2000 is compute-intensive."

Table 2.5: Difference in FOM between different image manipulations (including M-NONE) together with the 95 percent confidence intervals (CIs). Marked in gray are cells with a CI that does not include 0, which implies a statistically significant difference in FOMs ($p < 0.05$).

Compared manipulations			Difference in FOM	95% CI
M-NONE	vs	M-Blur	-0.0704	-0.2202, 0.0795
M-NONE	vs	M-Gamma	-0.1204	-0.2702, 0.0295
M-NONE	vs	M-ColSat	0.0389	-0.1109, 0.1887
M-NONE	vs	M-Noise	-0.0759	-0.2258, 0.0739
M-NONE	vs	M-JPG	-0.2037	-0.3535, -0.0539
M-Blur	vs	M-Gamma	0.0500	-0.0998, 0.1998
M-Blur	vs	M-ColSat	-0.1093	-0.2591, 0.0406
M-Blur	vs	M-Noise	0.0056	-0.1443, 0.1554
M-Blur	vs	M-JPG	0.1333	-0.0165, 0.2832
M-Gamma	vs	M-ColSat	-0.1593	-0.3091, -0.0094
M-Gamma	vs	M-Noise	-0.0444	-0.1943, 0.1054
M-Gamma	vs	M-JPG	0.0833	-0.0665, 0.2332
M-ColSat	vs	M-Noise	0.1148	-0.0350, 0.2646
M-ColSat	vs	M-JPG	0.2426	0.0928, 0.3924
M-Noise	vs	M-JPG	0.1278	-0.0221, 0.2776

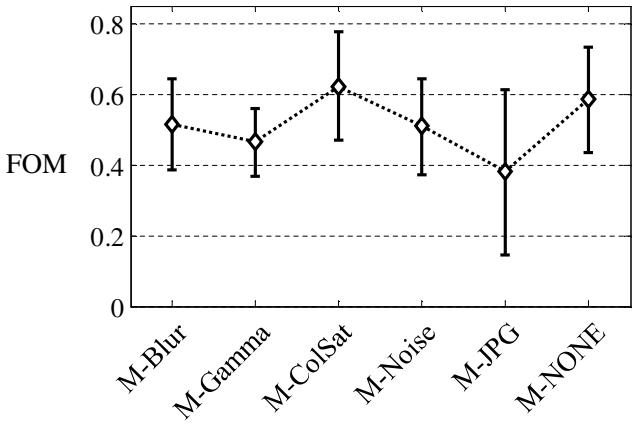


Figure 2.6: JAFROC FOM for each considered type of image manipulation (including M-NONE). The FOM is averaged over observers and the error bars correspond to 95 percent CI.

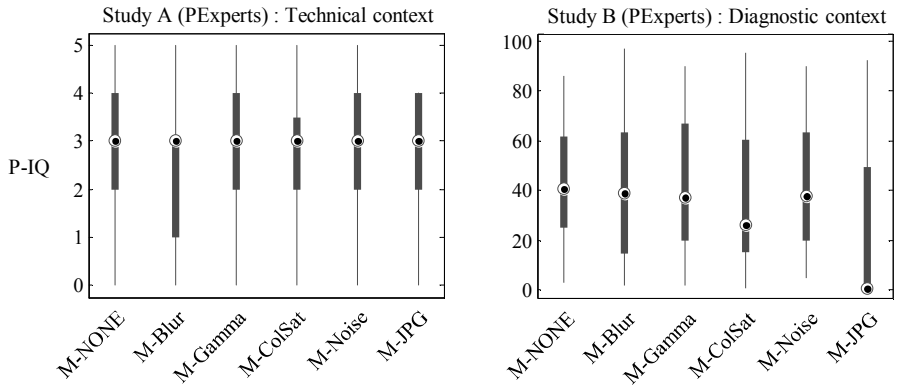


Figure 2.7: Overall IQ (P-IQ) ratings by PExpert observer group: (left) results from Study A (pooled across all Tiss1, Tiss2, and Tiss3), discrete 6-point rating scale from 0 to 5; (right) results from Study B, continuous rating scale from 0-100 percent. For both studies, a higher rating score corresponds to higher P-IQ. In both plots, the x-axis represents the type of image manipulation (M-NONE, M-Blur, M-Gamma, M-ColSat, M-Noise, M-JPG). Each box in the plot indicates the median, the IQR, the 1.5 IQR interval (whiskers); no "outliers" (measured points outside of the whisker range) have been identified.

2.5.4.2 TechIQ. Perceived overall IQ

The right plot in Figure 2.7 describes the P-IQ ratings from Study B. According to the Kruskal-Wallis test, PExperts found JPG compressed images to be of significantly lower quality compared to the reference images (as well as compared to any other category of manipulation). Note that this outcome agrees with the preceding JAFROC analysis which suggested a significantly lower diagnostic performance for M-JPG compared to the M-NONE images. Thus, in this particular case where the TechIQ and the TaskIQ were assessed "in parallel" (under the same experimental context), the two concepts of IQA seem to be in agreement for how they rank the manipulations. However, the relationship between TechIQ and TaskIQ may not always be like that; this issue is discussed further in Section 2.6 where we compare the two studies in more detail.

2.6 General discussion

After the detailed presentation of each technical and task-based human study, respectively in Section 2.4 and Section 2.5, we now compare the two study setups for their outcomes and discuss the possible causes and consequences of the observed agreements and disagreements. We focus on the PExperts exclusively as those observers all took part in both studies.

2.6.1 Does beauty mean utility?

Firstly, we compare the primary outcomes of the two study designs focusing on the PExpert observer group. On the one hand, Study A suggested no significant differences between P-IQ ratings of manipulated and corresponding reference images. Thus, according to the "beauty" of images, all manipulated images were of the same quality as the corresponding unmanipulated (reference) images. On the other hand, the JAFROC analysis from Study B (see Table 2.5) suggested that the diagnostic performance had dropped significantly for M-JPG images compared to the M-NONE images. Therefore, according to the "utility" of images, the JPG compressed images were of lower quality compared to the unmanipulated images. Evidently, Study A and Study B lead us to different findings.

Importantly, the disagreement between the two studies suggests that - in the case where we aim to evaluate the usefulness of the images for a specific task, *i.e.* the TaskIQ of the images - it might be misleading to rely on the TechIQ of images. That is, assessing the TaskIQ of images appears to require the task-based approach to IQA; TechIQ may not mean TaskIQ.

Intriguingly, though, in Section 2.5.4 we observed that the quality ranking suggested by P-IQ ratings collected during Study B were in agreement with those suggested by the diagnostic performance estimates from the same study. This is discussed next.

2.6.2 Context of the experiment: How does it matter?

For the purpose of this analysis, P-IQ results from Study A and Study B are reviewed in Figure 2.7. By comparing the P-IQ ratings from Study A (TechIQ) and Study B (TaskIQ), respectively the plot in the left and the plot in the right of Figure 2.7, we note the difference between the suggested ranking of the M-Any images relative to the M-NONE images: Study A suggests no differences, while Study B suggests that the M-JPG are of significantly lower quality than the M-NONE. Thus, it appears from our results that the human answer to the question of "How would you judge the overall quality of the image?" is not always the same, but rather the answer (the assigned IQ rating) may be influenced by the *context* in which the question is asked (explained shortly).

Of course, the variation in the ratings of a single individual when judging the same image on multiple occasions (reproducibility) could be attributed to the well-known and much studied effect of the "intra-reader" (intra-observer) variability. However, the effect of a single observer can hardly explain the difference in the average performance of multiple observers which occurs in our study (though we keep in mind that the number of observers in our study is limited and prevents us from making any strong conclusions).

More likely, the other two factors could be contributing to this effect: the instructions and the context. First, we recall that in Study B the PExperts were instructed to

judge the overall quality of the images according to their own personal criteria, while in Study A the observers received some training about the attributes of IQ (blur, noise, contrast, color saturation). Therefore, it is possible that the training they received in Study A led the PExperts to focus on some specific types of image impairments (artifacts) but because they were otherwise little familiar with them, the task might have been distracting and/or confusing. However, from the range of the rating scales which the PExperts used for P-IQ (see Figure 2.7) it seems that the PExperts were more confident of their P-IQ judgments in Study A (P-IQ ratings fall towards the mid and upper part of the scale) compared to Study B (P-IQ ratings were largely in the lower half of the scale). Of course, we note also that the two scales differ in their nature - discrete in Study A versus continuous in Study B. This in itself could be the topic for further discussion [Winkler, 2009], however, it is beyond the scope of our considerations here.

The second factor contributing to the disagreement between the P-IQ outcomes of the two study setups could be due to the context in which the observer is asked to judge the IQ. In Study A, the observer's task is exclusively about the quality of the images, and except for the image content itself (pathology tissues), there is no mention of the clinical context whatsoever; we refer to this context as "technical". In contrast, in Study B, the observers are asked to perform a clinically relevant task in addition to evaluating the images for their quality. Thus, this experiment has a very obvious "clinical" context. Interestingly, the end results for TechIQ and TaskIQ obtained in the clinical context are in full agreement (both TechIQ and TaskIQ measurements suggested the drop in quality from M-NONE to M-JPG images) while those from the technical context are not (TechIQ measurements from the technical context failed to detect the drop in quality of M-JPG). So could it be that the *context* (created by the actual clinical task) is affecting ("helping") the *overall IQ* judgment? In light of the latest findings of the significant differences in viewing behaviour (gaze response) in the task of rating quality versus free-looking at an image (no goal/task specified), the factor of experimental context could be a perfectly valid candidate for future investigation. It could be that the context of the experiment deserves more attention than it has been granted so far.

For example, when collecting human data for the purpose of developing numerical IQ measures, it might be of interest to tie the questions about different quality attributes to the task of interest, perhaps by asking the observers to perform the clinical task and evaluate the quality of the data within the same context, similar to what was done in our experiment from Study B. In current research practice, these two types of questions - clinical task versus quality ratings - are commonly asked in different contexts (during separate experimental sessions). In fact, most often, only one type of question is considered, either a clinical task or the P-IQ-attribute ratings.

2.7 Conclusion

In this chapter, we reported about two observer studies conducted for the purpose of assessing the impact of various image manipulations on the quality of digital veterinary pathology slides under the H&E staining. Study A measured perceptual IQ (P-IQ-attribute) ratings by three groups of observers (grouped by expertise profile): PExperts, PStudents and IExperts. Study B evaluated the performance of the PExperts in the task of detection and localization (FROC study) of inclusion bodies in the exact same set of images. In addition, Study B collected the ratings of overall IQ (P-IQ); thus TechIQ was also evaluated in the clinical context of Study B. The studies were preliminary and aimed at guiding future more in depth research of the topic. Accordingly, the size of the experiments was limited and no strong conclusions were made. Rather, the results were used as a discussion aid for some important issues concerning the TechIQ versus the TaskIQ approach to the evaluation of medical images.

To our best knowledge, no reports in the literature investigate the relation between P-IQ-attributes for the case of digital pathology nor do they compare P-IQ-attribute ratings between medical and imaging experts. Overall, our analysis in Study A points to blur and noise in the images (for IExperts) and changes in color saturation and gamma (for PExperts) to have more influence on the P-IQ-attribute ratings than other considered types of image manipulation. The effects of JPG compression were found disturbing by IExperts but not by PExperts. Furthermore, our data suggests that the initial criteria for judgment of P-IQ-attributes in digital pathology images is different for subjects of different expertise profiles, especially for PExperts and IExperts, while PStudents appear more conservative.

The observed discord between different observer groups advises, on the one hand, against guiding the development of pathology specific image algorithms or imaging systems by psychovisual TechIQ responses of non-expert human observers. For one thing, that would prevent the risk of inappropriate optimization of image parameters. Instead, the images should be optimized based on the preferences of the their actual users, the pathology experts. On the other hand, it has to be determined if the TechIQ of images, even when judged by the expert image users, can accurately predict the TaskIQ of images.

This concern has been addressed through another novel aspect of this work – the comparison between two approaches to IQA: TechIQ (Study A) versus TaskIQ (Study B). Based on the results from Study B, the diagnostic performance was affected only by compression, whereas other considered image manipulations had no significant effect. The same was suggested by the P-IQ ratings from Study B. Thus, within this study – the experimental context which involved a clinical task, the two quality concepts TechIQ and TaskIQ were in agreement. Interestingly, however, TechIQ findings that came from Study A – the purely technical context, were suggestive of a conflict between TechIQ and TaskIQ. Based on these observations, and considering evidence from existing literature, the TaskIQ approach definitely comes forward as the pre-

ferred, more reliable strategy for human IQA of medical images.

The contributions reported in this chapter resulted in one international conference proceedings [Platiša et al., 2013b] and international conference abstract [Platiša et al., 2013a]. As a co-author, one conference proceedings is accepted for publication [Kumcu et al., 2014] and three conference abstracts were published, all concerning the observer studies from Section 2.2.1: [Kumcu et al., 2013] investigating the effects of latency in laparoscopic video and [Kumcu et al., 2011a] and [Kumcu et al., 2011b] evaluating the quality of compressed laparoscopic sequences.

3

Models for task-based quality assessment of volumetric images

We advocate in Chapter 2 that medical image quality (IQ) should be assessed in terms of how useful images are for a specific clinical task – the task-based IQ (TaskIQ), rather than in terms of how excellent are specific image features – the technical IQ (TechIQ). In this chapter, we address the problem of TaskIQ assessment for the task of detecting lesions in volumetric (3D) medical images, which has been little explored so far. We study mathematical models (*model observers*) for estimating detection performance (either human-like or information-based, see Figure 1.2) assuming the lesion is *exactly known* (predefined shape, size, location). First, we provide an overview of the state-of-the-art and review the general strategies for the treatment of the 3D image data in the context of detection tasks. Then, we propose two novel designs of a 3D model observer. Finally, we conduct a simulation study to evaluate and discuss the pros and cons of the different model designs (three found in literature and ours two).

3.1 Introduction

Today, medical imaging is an essential part of clinical practice. The primary goal of medical imaging is to assist physicians in the diagnostic process. Given the seriousness of a diagnostic error, reliable and valid IQ assessment (IQA) is of fundamental importance in optimization and evaluation of medical imaging systems.

In its most general sense, IQ is often characterized as a measurement of image impairment. To that end, a number of "task-independent" (technical IQ, TechIQ) measures have been defined to evaluate a great range of factors which may affect the quality of a medical image: noise [National Electrical Manufacturers Association (NEMA), 2007], contrast resolution [American Association of Physicists in Medicine, 1993], and spatial resolution [Lodge et al., 2009], to mention just a few.

However, medical images are inherently *task-specific* rather than task-independent. In this respect, IQ for medical applications shall be defined in terms of how well, given the images, the specific diagnostic task can be performed by a physician [Judy et al., 1981, Myers et al., 1986]. In that manner, the task-based IQA is determined by the following four factors [Barrett, 1990]:

- the task of interest,
- the image data,
- the observer to perform the task, and
- the measure of observer performance.

In general, the diagnostic task in medical imaging is one of the following three: *estimation*, quantifying one or more parameters of interest using the given image data; *classification*, deciding to what class an image belongs; or hybrid *estimation-classification*, when estimation and classification are combined [Barrett and Myers, 2004]. In our work, we focus on one particular classification task called *signal detection* in which the image is classified either as signal-absent (normal clinical case) or as signal-present (abnormal clinical case). Tumor detection in PET scans, bone metastasis detection in bone SPECT scans, and mass detection in breast tomosynthesis are some common examples of relevant clinical tasks.

Until recently, medical images were limited to single-slice or 2D views, often projections or reconstructed 2D images. Thus, the detection tasks concerned planar signals in 2D images. In recent years, the advent of volumetric image acquisition and visualization (PET/SPECT, MRI, breast tomosynthesis, CT) has profoundly shifted the paradigm towards the detection of signals using multi-slice reconstructed image data [Reiner et al., 2001, Andersson et al., 2008, Rahmim and Zaidi, 2008]. Following these trends, assessing and optimizing IQ for 3D image analysis is one of the major challenges in medical imaging today.

The most obvious and currently still the most widely used task-based assessment of medical IQ is a human observer study. In such studies, the observers (subjects) are often true medical experts asked to make a diagnostic decision for the test images, either synthetic or real clinical ones. To their disadvantage, human observer studies are often time consuming and expensive. As an alternative, mathematical model observers may be used [Barrett et al., 1993, Barrett et al., 1995]. In general, two major types of model observers can be identified [Barrett and Myers, 2004]: *ideal observers* which estimate an upper bound on the signal-detection performance of any observer [Barrett et al., 1995, Park et al., 2003, Kupinski et al., 2003, Gallas and Barrett, 2003], and *anthropomorphic observers* which are designed to mimic human observer mechanisms and performance in a given detection task [Eckstein et al., 1998, Abbey and Barrett, 2001]. Commonly, two figures of merit are used to quantify observer performance in a binary classification task [Swets and Pickett, 1982, Metz, 1993, Barrett, 1990, Barrett

et al., 1998]: the area under the receiver operating characteristic curve (area under the ROC curve, AUC) and the detection signal-to-noise ratio (SNR).[1]

In signal detection theory [Green and Swets, 1966], the observer which has a full knowledge of the statistical information of the image data is known as the Bayesian ideal observer (IO). The IO is optimal among all observers, either human observer or model observer, in the sense that it maximizes diagnostic accuracy as measured by the AUC. Consequently, for design and optimization of data acquisition hardware, detection performance of the IO is preferred over any other observer. In practice, however, it is often difficult, if not impossible, to derive or estimate the IO performance. This is due to the high dimension and great complexity of the image statistics that are unknown and poorly estimated for real clinical data sets. The IO is tractable only for simple stylized settings, such as when the data is Gaussian, in which case the IO is linear.

It needs no debate that the clinical detection task is a complex mechanism to model, already in 2D, let alone in 3D, and thus simplifications are inevitable. This concerns both the observer model and the image data. One practical alternative to the IO is the ideal *linear* observer known as the Hotelling observer (HO). The HO is optimal among all linear observers in that it maximizes the SNR [Barrett and Myers, 2004]. Additionally, when the image data are Gaussian distributed, the HO is equal to the IO. Another simplification for the observer models is the so-called *channelized Hotelling observer* (CHO) proposed by Myers and Barrett [Myers and Barrett, 1987]. In essence, the CHO is an HO constrained to the "channelized" (filtered) image data. Originally, the channels were inspired by the properties of human visual system (HVS) related to examination of the data through frequency selective channels. An important advantage of the channelized models over the non-channelized ones is the dimensionality reduction of the problem, which has been discussed by [Barrett et al., 2001].

Depending on the properties of the channels relative to the image statistics in the task, the CHO can be used either to approximate the IO (*efficient* channels) or to track humans (*anthropomorphic* channels). For example, in 2D images, [Gallas and Barrett, 2003] found Laguerre-Gauss (LG) channels to be efficient in detection tasks using various lumpy backgrounds and rotationally symmetric signals. Not limited to types of backgrounds and signals are the singular vector channels of the system's singular-value-decomposition (SVD) used by [Park et al., 2009b, Park and Clarkson, 2009] which only require the system to be linear and the system's response functions to be known. Most recently, [Witten et al., 2009] investigated channels chosen by the partial least squares (PLS) method, which identifies channels based on the image and truth data covariance. Regarding anthropomorphic channels, their most common feature is that they have low or no response to low-frequency data, such as Gabor filters used in

[1] Note that the definition of the SNR used in the context of task-based IQA (in this book, Chapter 3 and Chapter 4) differs from the one commonly encountered in electrical engineering – the ratio of the signal power to the noise power. Further information can be found in Section 3.2.3.

the study by [Eckstein et al., 1998] or the difference-of-Gaussian (DOG) and square channels which [Abbey and Barrett, 2001] used in their experiments.

As an attempt to allow (approximate) computation of the IO also for the images from real life problems, whose statistics are often complex or unknown, the channelized IO (CIO) has been proposed. The authors demonstrate that the CIO using LG channels [Park et al., 2006, Park et al., 2007b], or more generally using system singular vectors [Park et al., 2009b, Park and Clarkson, 2009] or PLS channels [Witten et al., 2009], could well approximate the IO even for non-Gaussian images.

The most simplified approach of task-based IQA restricts the task of interest to detecting whether a known object (signal) is present at one specified location in a known background, the so-called binary signal-known-exactly and background-known-exactly (SKE/BKE) detection task [Myers et al., 1985, Myers and Barrett, 1987, Kim et al., 2004]. More complicated and more clinically relevant are the paradigms of background-known-statistically (BKS) [Rolland and Barrett, 1992,Burgess et al., 2001, Abbey and Barrett, 2001,Gallas and Barrett, 2003,Park et al., 2003,Park et al., 2007c, Park et al., 2007b,Chen et al., 2002,Lartizien et al., 2004, Young et al., 2009, Gifford et al., 2005, Park et al., 2009a] and signal-known-statistically (SKS) [Gifford et al., 2005, Park et al., 2005, Castella et al., 2009, Goossens et al., 2010, Zhang et al., 2013] which incorporate background and signal variability, respectively. For the scope of this work, we focus on SKE/BKS tasks.

In recent publications, several authors proposed different approaches for treating the 3D image data during the process of signal detection. The most direct way to migrate the model observer for 2D detection task to the 3D detection task is to use a conventional 2D (planar) CHO and apply it on a single image slice only, the slice where the signal is centered (mostly concentrated). We refer to this approach as *single-slice* CHO (ssCHO). It has been used by [Liang et al., 2008], for example, to estimate observer performance in sequence-browsing mode of volumetric image reading. As the authors pointed out, the limitation of the ssCHO is that model observers which are designed for use in pure 2D detection tasks do not incorporate information about signal contrast in the z-direction nor the spatial correlation of the background and signal in the adjacent slices.

A similar motivation underlies the analysis by [Kim et al., 2004] who compared the behavior of 2D and 3D implementations of the numerical observers for simulated whole-body PET oncology imaging. Their results indicate that there is a significant increase in SNR or detectability of volumetric model observers relative to planar ones. Similarly, [Lartizien et al., 2004] used 3D implementations of model observers with 3D channels to compare different acquisition protocols in whole-body PET imaging, and found these to be a useful tool for their task of interest. We call a 3D implementation of the CHO a *volumetric* CHO (vCHO).

[Chen et al., 2002] proposed a more sophisticated two-layer model which combines 2D CHOs followed by an HO. The model which they called a multi-slice CHO-HO was used to process simulated multi-slice multi-view images similar to SPECT

myocardial perfusion scans. First, the image slices of each of the three orthogonal views (coronal, sagittal and axial) were channelized and the 2D CHO was computed for each slice and each view, giving arrays of the decision variables. Then, an HO was applied on these decision variable arrays to obtain a single scalar detection score for the 3D image, known in statistical hypothesis testing as the *test statistic*. This approach was guided by the assumption that, for multi-slice images, human observers make their detection decision in a two-stage process. The first stage assessing each slice separately and the second stage integrating these slice assessments to yield the final classification decision. Later, [Gifford et al., 2005] tested two different processes for modeling the observer capacity for integrating the information from multiple slices in the image sequence. One process describes an observer that is able to integrate the slice information by computing the sum of the decision variables for each slice to represent the final test statistic for the image sequence. The other process supposes that the observer is unable to do any integration, and instead the image test statistic is assigned the maximum value of the decision variables across slices. Further on, we use the term *multi-slice* CHO (msCHO) to refer to any approach which treats the 3D image as a conglomerate of multiple slices rather than just a single volume.

Most recently, [Young et al., 2009] used 2D projections of 3D breast tomosynthesis data to approximate the performance of ideal linear observer. Unlike the conventional CHO that would use a single 2D projection only, they built a CHO model that uses concatenated channelized angular projections. By doing so, Young *et al.* were able to incorporate correlations between multi-projections. Again, their preliminary results indicate that the observer using multiple projections outperforms the single-slice observer in their considered range of image acquisition parameters.

The aim of our work is to identify the candidate model observers for the treatment of 3D images and to evaluate their performance with respect to a range of parameters which could be of importance for the practical applications, such as image signal properties (size, amplitude), image background properties (structure, correlation) and size of the image data sample (number of training and test images). Ultimately, these investigations would serve as a basis for building the anthropomorphic models for 3D images. In that respect, our investigations assume a sequential three-stage approach to modeling human performance.

Stage 1. Select the candidate models.

Stage 2. Compare candidate model predictions to human performance results on actual classification tasks.

Stage 3. Modify the best candidate model(s) to better predict human performance.

Given the scarce previous research on the topic, this thesis begins with the study of the candidate models, reported in this chapter. In particular, we consider the following CHO models: the ssCHO [Myers and Barrett, 1987, Gallas and Barrett, 2003], three msCHO designs, and the vCHO model [Kim et al., 2004, Lartizien et al., 2004].

The three multi-slice designs include the model proposed by [Chen et al., 2002] only restricted to a single view (either coronal, sagittal or axial), and two novel msCHO models introduced in this chapter: one guided by the assumptions from the work of [Chen et al., 2002] and one inspired by the recent work of [Young et al., 2009]. The models are evaluated in a series of multiple-reader multiple-case (MRMC) experiments (relying on the training and testing paradigm). To account for potential influence of the background structure, we analyze the models within four different data setups, all SKE/BKS: white Gaussian noise (WNB), correlated Gaussian noise (CNB), lumpy backgrounds (LB) [Rolland and Barrett, 1992] and clustered lumpy backgrounds (CLB) [Bochud et al., 1999, Liang et al., 2008]. Especially, for the two Gaussian data setups we also estimate the IO strategy to serve as a point of reference in evaluating the range of disparity among the CHO models.

Overall, our results show that the volumetric model outperforms the others in all four setups. The multi-slice models are the next best, and the single-slice model expectedly achieves the lowest detection scores. At the same time, the disparity between the models is most notable for most difficult detection tasks (*e.g.* detecting a Gaussian signal in a correlated Gaussian noise background when their parameters are very similar) and it gets less pronounced as the difficulty of the task drops (*e.g.* detecting a Gaussian signal in a white noise background).

Further contributions of this thesis towards modeling human performance in volumetric detection tasks relate to the stage two of the aforementioned three-stage development process and they are elaborated in Chapter 4.[2] Lastly, concerning stage three of the process, we refer to the most recent literature report by [Michielsen et al., 2013] and that by [Avanaki et al., 2013]. They use the msCHO models proposed in this chapter as the basis for their CHO designs which successfully predict human performance in detection of lesions (masses and micro-calcifications, respectively) in digital breast tomosynthesis (DBT) images.[3]

The research related to the model observers has been performed within the framework of the "Medical Virtual Imaging Chain" (MEVIC) project financially supported by iMinds. The project involved collaboration with multiple academic and industrial

[2] Specifically, in Section 4.4 we study the task of detecting mass lesions in digital breast tomosynthesis (DBT) images and compare the msCHO performance to that of the human performance measured in the study by [Marchessoux et al., 2011]. Moreover, we conduct a human observer study which explores detection performance trends of humans under different image presentation modes: single-slice (planar) versus multi-slice sequence-browsing image presentation. This report can be found in Section 4.6.

[3] In particular, [Avanaki et al., 2013] were interested in modeling humans under different browsing speeds. They modified our msCHO model by incorporating the spatio-temporal contrast sensitivity function (CSF) of the HVS and compared it to the human performance measured earlier by [Diaz et al., 2011]. The extended model was able to predict the detectability trends of humans. On the other hand, [Michielsen et al., 2013] used the msCHO to evaluate the performance of newly developed reconstruction algorithms on the task of detecting micro-calcifications in DBT images. Their modification to the msCHO model concerned the channels. In place of the LG channels from our study, [Michielsen et al., 2013] created the channels by applying the inverse Fourier transform to an elliptical band in the frequency domain chosen to approximately match the non-isotropic point spread function of the DBT images; the exact shape of the channels was of little influence on the msCHO. The authors found high correlation between human and model performance.

partners including Dr. Cédric Marchessoux and Dr. Tom Kimpe (Barco N.V., Belgium). In addition, we closely collaborated on these topics with Dr. Aldo Badano, Dr. Brandon D. Gallas, and Dr. Subok Park, (U.S. Food and Drug Administration, USA), Prof. Bart Goossens and Dr. Ewout Vansteenkiste (Department of Telecommunications and Information Processing, Ghent University, Belgium).

The chapter is organized as follows. Section 3.2 describes the models of image objects used in the study and provides the essential background information about the model observers. In Section 3.3 we review the existing CHO designs for 3D images (ssCHO, vCHO, multi-slice CHO-HO) and introduce the two novel msCHO models. These five models are considered potential candidates for anthropomorphic models. Our experimental study is explained in Section 3.4 and the results are presented and discussed in Section 3.5. The discussion also includes some practical considerations about the choice of the model and the selection of model parameters in different types of applications (*e.g.* evaluation of an imaging system in early versus in final development phase, or evaluation of images with few-slice signals versus images with many-slice signals). Lastly, Section 3.6 draws some conclusions from this work.

3.2 Mathematical background

We are interested in a binary classification task determined by two hypotheses: signal is absent (H_0) or signal is present (H_1). An observer decides which of these two is true for a given image denoted by a column vector $\mathbf{g}$. The entries $g_m, m = 1, ..., M$, are the intensity of image pixels in 2D data or image voxels in 3D data, and M is the number of elements (pixels or voxels) in the image. An observer is defined by its *discriminant function* which maps an image $\mathbf{g}$ to its test statistic, $t = t(\mathbf{g})$. The decision is made by comparing the test statistic to a certain threshold, t_0. When t is greater than t_0, the signal is considered detected, hence H_1 holds, and the image is classified as signal-present. Otherwise, H_0 is satisfied and the image is classified as signal-absent.

In the remaining of this section, we introduce the image models considered in our study, briefly outline the fundamentals of the Bayesian ideal observer. and review the mathematical framework for the linear observer models.

3.2.1 Object models

Since our objective is to investigate fundamental aspects of observer models for multi-slice images, we consider three-dimensional images with known statistical properties and different levels of complexity. This provides a controllable test environment and allows for automated generation of a large number of random realizations which increases the statistical significance of the experimental results.

Let us denote $\mathbf{s}$ the signal to be detected, $\mathbf{b}$ the noiseless image background and $\mathbf{n}$ the measurement of noise in the image. Then the data under the two hypotheses are

Table 3.1: Image data parameters. The following notation applies (see text for details): M – number of voxels in the image; M_{FOV} – number of voxels in the field of view (LB, CLB); σ_s – spread parameter of the 3D Gaussian signal; a_s – magnitude of the 3D Gaussian signal; σ_b – standard deviation of the 3D Gaussian kernel (CNB); or spread parameter of the 3D Gaussian lump (LB); a_b – peak intensity level in the background image; $\overline{K}$ – mean number of lumps in the field of view (LB, CLB); L_x, L_y and L_z – characteristic lengths of asymmetrical lumps in x, y and z directions respectively (CLB).

Background category	Background 3D image, $\mathbf{b}$	Gaussian 3D signal, $\mathbf{s}$
White noise (WNB)	$\sim N(0,1)$ $M = 64^3$	$\sigma_s = 8$, $a_s = \{0.015, 0.025, 0.035, 0.045\}$
Colored noise (CNB)	$\sigma_b = 8$, $M = 64^3$	$\sigma_{s1} = 8$, $a_{s1} = \{0.25, 0.5, 0.75, 1\}$ $\sigma_{s2} = 5$, $a_{s2} = \{0.01, 0.015, 0.02, 0.025\}$ $\sigma_{s3} = 3$, $a_{s3} = \{0.0025, 0.0035, 0.0045, 0.0055\}$
Lumpy background (LB)	$\sigma_b = 8$, $a_b = 255$, $M = 64^3$ $M_{\mathrm{FOV}} = 128^3$, $\overline{K} = 800$	$\sigma_s = 8$, $a_s = \{4, 8, 12, 16\}$
Clustered lumpy background (CLB)	$L_x = 3$, $L_y = 2$, $L_z = 3$, $a_b = 255$, $M = 64^3$, $M_{\mathrm{FOV}} = 128^3$, $\overline{K} = 80$, $\overline{N} = 20$	$\sigma_s = 8$, $a_s = \{4, 8, 12, 16\}$

given by

$$H_0 : \mathbf{g} = \mathbf{b} + \mathbf{n}, \tag{3.1}$$

$$H_1 : \mathbf{g} = \mathbf{b} + \mathbf{s} + \mathbf{n}. \tag{3.2}$$

In our case, four different models are considered for $\mathbf{b}$ while the model of $\mathbf{s}$ is kept the same for all four background models. The amount of measurement additive white Gaussian noise $\mathbf{n} \sim N(0, \sigma_n)$ is small and not disturbing the statistical properties of the background. The models we use for background and signal simulations are described in the remaining of this subsection, and their parameters are summarized in Table 3.1.

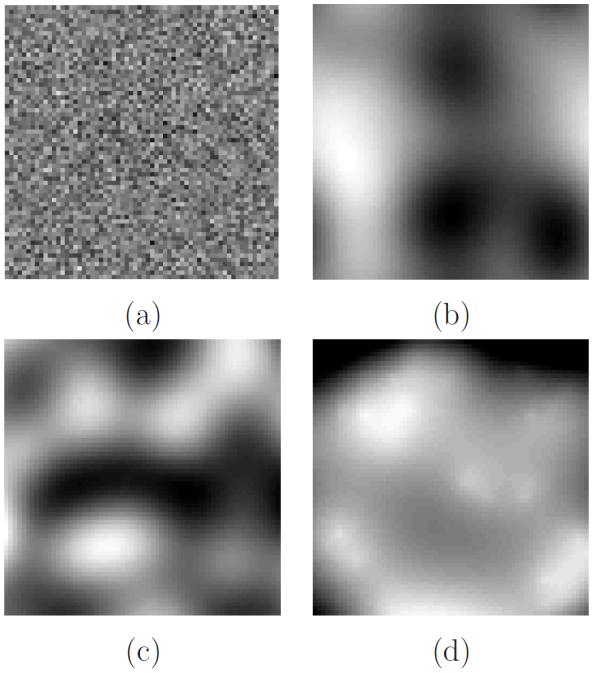


Figure 3.1: Four image categories: (a) white noise background (WNB), (b) correlated Gaiussian noise background (CNB), (c) lumpy background (LB), and (d) clustered lumpy background (CLB). In each case, a randomly selected slice from the image volume is presented. Detailed parameters of the background images are given in Table 3.1.

3.2.1.1 Image backgrounds (BKS)

The criteria for choosing the background models are twofold. On the one hand, given that the purpose of the model observers is assessment of medical images, we aim at image data models which may be of clinical relevance. On the other hand, we are interested in estimating the IO strategy for the selected data as a point of reference for comparing the CHO models. In most cases, these two criteria exclude each other: the IO performance is very difficult to estimate for clinical images because their statistics are understandably complex and often unknown.

In order to keep the analysis general, we select the following four different categories of background images: white Gaussian noise (WNB), correlated Gaussian noise, or colored noise backgrounds (CNB), lumpy backgrounds (LB) [Rolland and Barrett, 1992], and clustered lumpy backgrounds (CLB) [Bochud et al., 1999, Liang et al., 2008]. Example background images with added measurement noise are shown in Figure 3.1. These correspond to signal-absent images in the study.

The first two models assume Gaussian statistics so that the Bayesian IO strategy is

readily calculable for these problems [Barrett and Myers, 2004]. We will use these IO calculations to evaluate the non-ideal model observers (variants of CHO) against the theoretical upper bounds of the performance, to be explained in Section 3.2.2. In contrast, the LB and CLB models are used as representatives of non-Gaussian data. The two-dimensional CLB have been shown by [Bochud et al., 1999] to have a close visual appearance to real mammographic backgrounds. Recently, [Castella et al., 2008] used a genetic algorithm to optimize the CLB generation and achieve even more realistic mammographic texture synthesis. At the same time, both the LB and CLB models are statistically well described which allows automated generation of large ensembles of images required for the model observer experiments. Due to their complexity, the IO strategy for LB and CLB are not included in the present analysis. We remark here that for the two-dimensional non-Gaussian LB data [Kupinski et al., 2003] and [Park et al., 2003] have been able to estimate the IO and the channelized IO (CIO), respectively, using Markov-chain Monte Carlo (MCMC) techniques.

The simplest background we consider is the fixed background, $\mathbf{b}_{\mathrm{WNB}} = 0$. Since the statistics of these backgrounds are determined by the added measurement white noise, $\mathbf{n}$ (see Eqs. (3.1) and (3.2)), we refer to them as white noise backgrounds, WNB. Next in order of background complexity is CNB data, $\mathbf{b}_{\mathrm{CNB}}$, generated by convolving white noise with a 3D Gaussian kernel characterized by σ_b. The correlated Gaussian random backgrounds are sometimes also referred to as lumpy backgrounds, not to be confused with the LB as we use them in this study, which are non-Gaussian. We describe these next.

As defined by [Rolland and Barrett, 1992], a *lumpy background* $\mathbf{b}_{\mathrm{LB}}$ is produced by placing a random number K of lumps $l(\mathbf{r})$ at random locations $\mathbf{r}_k, k = 1, ..., K$ in the image. In our simulation, $\mathbf{b}_{\mathrm{LB}}$ is extracted from a larger field of view (FOV), $\mathbf{f}_{\mathrm{LB}}$, in order to avoid a boundary problem in generating the LB images. In particular, the size of $\mathbf{f}_{\mathrm{LB}}$ is $M_{\mathrm{FOV}} = 128^3$ voxels and the size of $\mathbf{b}_{\mathrm{LB}}$ is $M = 64^3$ voxels (see also Table 3.1). Formally, the LB images can be described as

$$\mathbf{f}_{\mathrm{LB}}(\mathbf{r}) = \sum_{k=1}^{K} l(\mathbf{r} - \mathbf{r}_k), \tag{3.3}$$

where $\mathbf{r}$ is a 3D vector denoting the spatial position and K is the number of lumps selected using a Poisson probability distribution with mean $\overline{K}$. For the LB images, the values of lump locations, $\mathbf{r}_k$, are selected using a uniform probability distribution over the support of the FOV, $\mathbf{f}_{\mathrm{LB}}$. The set of K lump locations may be referred to as a "lump map" of the image. We choose the lumps to be 3D Gaussian signals of magnitude a_b and with the spread parameter σ_b. By letting $|\mathbf{r}|$ denote the magnitude of the 3D vector $\mathbf{r}$, we can define the lump as

$$l(\mathbf{r}) = a_b \exp\left(\frac{-|\mathbf{r}|^2}{2{\sigma_b}^2}\right). \tag{3.4}$$

Finally, the most complex background we treat in this chapter is the CLB, denoted

$\mathbf{b}_{\mathrm{CLB}}$. The original concept of the two-dimensional CLB was introduced by [Bochud et al., 1999]. In [Liang et al., 2008], the 2D concept is extended to 3D with the assumption that the projection of a 3D CLB yields a 2D CLB with related characteristics in terms of the parameters of cluster and lump size and density.

As with the LB, to prevent potential boundary effects, the background $\mathbf{b}_{\mathrm{CLB}}$ of size $M = 64^3$ voxels is extracted from a larger FOV, $\mathbf{f}_{\mathrm{CLB}}$ of size $M_{\mathrm{FOV}} = 128^3$ voxels. The $\mathbf{f}_{\mathrm{CLB}}$ is created in a two step process. The first step is similar to the process with the LB, only now we shall refer to the "lump map" as the "cluster map" and use $\mathbf{r}_k, k = 1, \ldots, K$ to denote cluster (rather than lump) location. In the next step, each cluster position $\mathbf{r}_k$ is used as the spatial origin for placing a random number N_k of lumps. These N_k lumps are randomly positioned within the k-th cluster at locations $\mathbf{r}_{k,n}, n = 1, \ldots, N_k$. Thus,

$$\mathbf{f}_{\mathrm{CLB}}(\mathbf{r}) = \sum_{k=1}^{K} \sum_{n=1}^{N_k} l(\mathbf{r} - \mathbf{r}_k - \mathbf{r}_{k,n}), \tag{3.5}$$

where K stands for the number of clusters in the field of view $\mathbf{f}_{\mathrm{CLB}}$ and N_k is the number of lumps in the cluster k. Again similar to the LB, both K and N_k are selected using a Poisson probability distribution with mean values $\overline{K}$ and $\overline{N_k}$, respectively. The location of the k-th cluster, $\mathbf{r}_k$, is selected using a uniform probability distribution over the support of the $\mathbf{f}_{\mathrm{CLB}}$. To create CLB images, anisotropic 3D exponential blobs are used with characteristic lengths L_x, L_y and L_z in x, y and z directions, respectively. The details can be found in [Liang et al., 2008].

3.2.1.2 Spherically symmetric signal (SKE)

Signal-present images are created by adding the signal $\mathbf{s}$ to a background image $\mathbf{b}$. In particular, we use a spherically symmetric Gaussian blob created in 3D Cartesian space and centered in the image volume. Similar to the lump in LB backgrounds, the signal is defined by Eq. (3.4) only now we use a_s to represent signal magnitude and σ_s to denote signal spread parameter. The central slice from a sample signal volume is depicted in Figure 3.2(a) and the radial profile of the signal used in the study is given in Figure 3.2(b).

Parameters of both the backgrounds and the signals used in our experiments are listed in Table 3.1. The spread of the signal σ_s is chosen the same as as σ_b to correspond to a difficult detection task (which is of most interest for task-based IQA of today's advanced medical imaging systems), and the signal amplitudes a_s are chosen to cover the AUC range of approximately 0.6 to 0.9.

3.2.2 Observer models

According to the signal detection theory [Green and Swets, 1966], the observer is completely characterized by its discriminant function which assigns a scalar test statistic

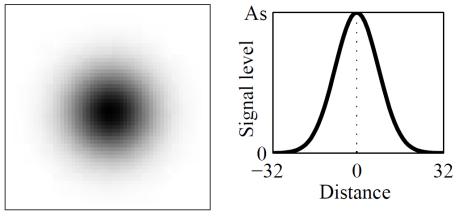


Figure 3.2: An example signal image. (a) central slice of the signal volume, size of the slice is 64×64 voxels, (b) contrast profile in the central slice of a simulated 3D Gaussian signal.

to each image object, $t = t(\mathbf{g})$. In the following, we introduce the ideal and the channelized mathematical model observers, and define their discriminant functions.

3.2.2.1 The ideal observer

The Bayesian ideal observer (IO) is defined as one that has full knowledge of the problem in terms of the conditional probability density functions of image data under each hypothesis, $\mathrm{pr}(\mathbf{g}|H_i)$, $i = \{1, 2\}$. Hence, the test statistic of the IO is defined as the likelihood ratio [Green and Swets, 1966],

$$\Lambda(\mathbf{g}) = \frac{\mathrm{pr}(\mathbf{g}|H_1)}{\mathrm{pr}(\mathbf{g}|H_0)}. \tag{3.6}$$

Clearly, calculation of the likelihood ratio, or more conveniently the log-likelihood ratio $\lambda(\mathbf{g}) = \ln \Lambda(\mathbf{g})$, requires knowledge of the probability density functions that make up the Eq. (3.6). In practical applications, these are often complicated or even unknown. Here, though, the IO can be derived for the WNB and CNB image models. The analytical expressions for calculating the SNR and AUC of the IO for these two image categories are given in Section 3.2.3.

For greater complexity of the image data statistics, analytical formulas for calculating the theoretical upper bounds for the observer performance cannot be derived. This applies even in cases of simulated data such as LB or CLB from our study and especially in cases of real clinical data. Rather, computation of the likelihood ratio in those cases requires specialized procedures to be developed. As we mention earlier, in current literature this has been done for 2D LB and CLB backgrounds using MCMC techniques [Kupinski et al., 2003, Park et al., 2003, Park and Clarkson, 2009].

3.2.2.2 Channelized observers

In case of real clinical images, it is often difficult or impossible to know the probabilities required to calculate λ. Primarily, this is caused by random variations in both anatomical background (bones, veins, organs) and the signal (size, shape and location

of the lesion) which are not all well understood to date and thus accurate models of those are not yet available. To circumvent this problem, a linear approximation of the IO has been defined, where linearity refers to the discriminant function

$$t(\mathbf{g}) = \sum_{m=1}^{M} w_m g_m. \tag{3.7}$$

Here, M is the number of elements in the image $\mathbf{g}$ and the weights w_m, $m = 1, ..., M$ form an image $\mathbf{w}$ called the *template* of the observer. Thus, the discriminant function may be written as a scalar product[4]

$$t(\mathbf{g}) = \mathbf{w}^t \mathbf{g}. \tag{3.8}$$

Commonly, the template $\mathbf{w}$ is estimated within the framework of linear discriminant analysis and the optimal linear discriminant is defined as the one which maximizes the SNR. In this context, the ideal *linear* observer is known as the Hotelling observer (HO) [Barrett and Myers, 2004]. First, let us denote by $\mathbf{K_g}$ the average of the ensemble covariance matrices of the signal-absent and signal-present data. It is defined as follows:

$$\mathbf{K_g} = \frac{1}{2}(\mathbf{K_{g,1}} + \mathbf{K_{g,2}}), \tag{3.9}$$

with $\mathbf{K_{g,i}} = \langle(\mathbf{g} - \overline{\mathbf{g_i}})(\mathbf{g} - \overline{\mathbf{g_i}})^t | H_i\rangle$, $i = \{1, 2\}$, and $\overline{\mathbf{g_i}} = \langle \mathbf{g} | H_i \rangle$. Then, the template of the HO is defined as

$$\mathbf{w_{HO}} = \mathbf{K_g}^{-1} \Delta\overline{\mathbf{g}}, \tag{3.10}$$

where $\Delta\overline{\mathbf{g}} = \langle \mathbf{g} | H_1 \rangle - \langle \mathbf{g} | H_0 \rangle$ and $\langle \cdot \rangle$ denotes ensemble average.[5] The template is often estimated from the images for which the ground truth is known a priori. We refer

[4] In the case of SKE/BKE tasks with additive white Gaussian noise (in our study, the WNB data), it can be shown (p.46–48 of [Gallas, 2001]) that the ideal template is the signal image, $\mathbf{w}_{\text{ideal}} = \mathbf{s}$. Then, the discriminant function can be written as $t(\mathbf{g}) = \mathbf{s}^t\mathbf{g}$ which resembles the expression for the well-known "matched filter" (or "correlation filter"). Note though that here it is a scalar product (the location of the signal is known), not a correlation. Due to the analogy, the ideal observer is often referred to as the *matched filter*.

[5] In the case of SKE tasks, the mean difference in the data under each hypothesis is exactly the signal image, *i.e.*, $\Delta\overline{\mathbf{g}} = \mathbf{s}$. Then, the right side of Eq. (3.10) can be rewritten as $\mathbf{K_g}^{-1}\mathbf{s}$. Further on, assuming the covariance matrix $\mathbf{K_g}$ is nonsingular, we can split $\mathbf{K_g}^{-1}$ into two pieces and write the former expression as $\left(\mathbf{K_g}^{-\frac{1}{2}}\mathbf{s}\right)^t \mathbf{K_g}^{-\frac{1}{2}}$. The transformation defined by the matrix square root of the matrix inverse of the covariance matrix $\mathbf{K_g}^{-\frac{1}{2}}$ is a process known as *whitening*, or *prewhitening* when it precedes other processing. In simple terms, the prewhitening means decorrelating the data. Having thus transformed the original Eq. (3.10) and given the remarks from footnote[4], we see that the ideal observer strategy is to first prewhiten the image data and then match it to the prewhitened signal. Therefore, this observer is often known as the *prewhitening matched filter* (PWMF). Based on the literature [Burgess et al., 1982, Myers et al., 1985], the detection performance of humans compared to that of the ideal observer seems a lot more similar for white noise images than for images with correlated noise – suggesting that the human visual system is unable to decorrelate the data. This motivated the approach of modeling humans with a matched filter but without the prewhitening filter. Such a model observer is known as the *non-prewhitening matched filter* (NPWMF) [Judy and Swensson, 1985].

to those as the *trainer data* and to the related process as the *training phase*. Next, in the *testing phase*, the estimated observer template is used to classify the images for which the ground truth is unknown, the *tester data*.

When the images are Gaussian random vectors, the HO equals the IO [Barrett and Myers, 2004]. However, to their disadvantage, both the IO and HO encounter the difficulty of high-dimensionality computations [Barrett et al., 2001]. The main difficulty in computing the HO stems from the inversion of a large covariance matrix, $\mathbf{K_g}$, which is used in Eq. (3.10) to estimate the observer template, $\mathbf{w}_{\mathrm{HO}}$.

To overcome the dimensionality problem of the HO model, another variant of the linear observer named the *channelized* Hotelling observer was defined [Myers and Barrett, 1987]. The CHO may be seen as a specialization of the HO model which makes use of the frequency selective channels (inspired by the HVS) to reduce the dimensionality of the problem. The size of the channels is the same as that of the images (M-elements) and likewise they are represented as column vectors $\mathbf{u}_p$, $p = 1, ..., P$, where P is the number of channels (often around 10, or of that order).

In contrast to the HO where all image data is used to build the template $\mathbf{w}_{\mathrm{HO}}$, the CHO model only makes use of the channel outputs,

$$\mathbf{v} = \mathbf{U}^t \mathbf{g}, \tag{3.11}$$

where $\mathbf{U}$ denotes the channel matrix formed by concatenating the P channel vectors, $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_P]$. Knowing that the typical number of channels is of the order of 10, it is obvious that processing the images through channels greatly reduces the dimensionality of the problem, $P \ll M$.[6]

If we denote the ensemble covariance matrix of the channelized data as $\mathbf{K_v}$, the template of a CHO model is

$$\mathbf{w}_{\mathrm{CHO}} = {\mathbf{K_v}}^{-1} \Delta\overline{\mathbf{v}}, \tag{3.12}$$

where $\mathbf{K_v} = \mathbf{U}^t \mathbf{K_g} \mathbf{U}$ and $\Delta\overline{\mathbf{v}} = \mathbf{U}^t \Delta\overline{\mathbf{g}}$. Finally, the CHO test statistic is calculated as a linear combination of all channel responses, $t_{\mathrm{CHO}}(\mathbf{v}) = {\mathbf{w}_{\mathrm{CHO}}}^t \mathbf{v}$.

When selecting the channels for our experimental study in Section 3.4, our primary objective is achieving optimal performance of the models (rather than comparing their ability to mimic humans). Additionally, the following assumptions apply to our study: there is no preferred orientation in the correlation structure of the background and the signal is spherically symmetric positioned at a known location. Accordingly, we chose Laguerre-Gauss (LG) functions centered on the location of the signal; see Figure 3.3 for illustration. The details about the specific use of the channels in different CHO designs are given in Section 3.4.

[6] To illustrate this reduction, we refer to the parameter values from Table 3.1: the size of the image is $M = 64^3 = 262{,}144$ voxels while the number of channels P is of the order of 10. Note that real clinical images are usually much larger while the number of channels usually remains of the order of 10.

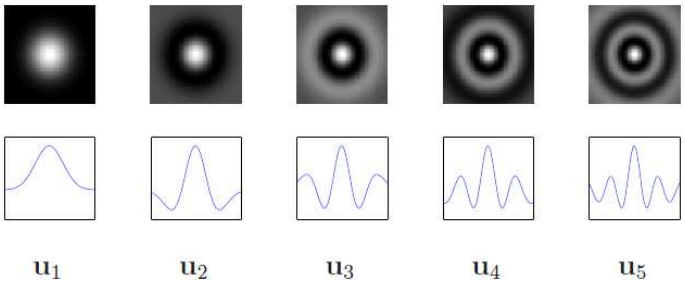


Figure 3.3: The first five LG channels with the spread parameter $a_u = 24$. Top: The images illustrate 2D channels or central slices of 3D channels. Bottom: Plots of the LG functions. For 3D channels, these plots are the same in planar view (xy-plane) as in the z-direction.

The LG channels are a product of Laguerre polynomials and Gaussian functions, and defined by

$$u_p(|\mathbf{r}|) = \frac{\sqrt{2}}{a_u} \exp\left(\frac{-\pi|\mathbf{r}|^2}{{a_u}^2}\right) L_p\left(\frac{2\pi|\mathbf{r}|^2}{{a_u}^2}\right), \tag{3.13}$$

where $|\mathbf{r}|$ denotes the magnitude of the spatial position, a_u is the spread parameter of the LG channel, and L_p denotes Laguerre polynomials defined by

$$L_p(x) = \sum_{k=0}^{p} (-1)^p \binom{p}{k} \frac{x^k}{k!}. \tag{3.14}$$

The weight of the polynomials is concentrated within a Gaussian envelope with spread σ_u, where $a_u^2 = 2\pi\sigma_u^2$. The procedure used for selecting the LG channel parameters in our study is described in Section 3.4.2 and the corresponding results are summarized in Table 3.4.

3.2.3 Performance measures

In the task-based IQA, AUC and SNR are often used to quantify the performance of the model observers [Barrett et al., 1998]. The SNR for binary classification tasks is defined on the test statistic $t(\mathbf{g})$ as

$$\mathrm{SNR}^2 = \frac{\langle t(\mathbf{g})|H_1\rangle - \langle t(\mathbf{g})|H_0\rangle}{\frac{1}{2}\mathrm{Var}(t(\mathbf{g})|H_1) + \frac{1}{2}\mathrm{Var}(t(\mathbf{g})|H_0)}, \tag{3.15}$$

where $\langle\cdot\rangle$ denotes the expected value and $\mathrm{Var}(\cdot)$ is the corresponding variance. For more detailed information interested readers are referred to the literature, *e.g.*, [Barrett et al., 1998, Barrett et al., 2006] and the references therein.

Alternatively, the detection SNR can be computed from AUC. In the first step, we use the image test statistics $t(\mathbf{g})|H_1$ and $t(\mathbf{g})|H_0$ and apply the Mann-Whitney-

Wilcoxon (MWW) statistic to estimate the AUC of a CHO model. Then, assuming the test statistics under both hypothesis are Gaussian, the relationship between SNR and AUC can be expressed as follows [Barrett et al., 1998]:

$$\mathrm{SNR} = 2\,\mathrm{erf}^{-1}(2\,\mathrm{AUC} - 1), \tag{3.16}$$

where $\mathrm{erf}(\cdot)$ represents the error function. If the two normally distributed test statistics also have the same variance, the SNR from Eq. (3.16) is termed the detectability index d' and it is commonly used for performance comparison in the domain of observer studies [Barrett and Myers, 2004]. We remark also that Eq. (3.16) shall be used only if $\mathrm{AUC} < 1$; otherwise, Eq. (3.15) applies.

For two image categories considered in the study, WNB and CNB, we can estimate the IO strategy and use it as a reference in evaluating the performance of the CHO models. For LB and CLB, those calculations shall not be included.

In the case of the IO, we first calculate the SNR and then use it in Eq. (3.16) to get the AUC. When the signal is assumed exactly known, the SNR of the IO equals

$$\mathrm{SNR}_\lambda = (\mathbf{s}^t \mathbf{K}^{-1} \mathbf{s})^{1/2}. \tag{3.17}$$

Here, $\mathbf{K}$ stands for the covariance matrix of the background: $\mathbf{K}_{\mathrm{WNB}} = {\sigma_{\mathrm{WNB}}}^2 \mathbf{I}$ in case of WNB, or $\mathbf{K}_{\mathrm{CNB}}(\mathbf{r}_i, \mathbf{r}_j) = {a_b}^2 (\pi {\sigma_b}^2)^{3/2} \exp\left(-(|\mathbf{r}_j| - |\mathbf{r}_i|)^2/(4{\sigma_b}^2)\right)$ in case of 3D CNB for which the Gaussian kernel is determined by Eq. (3.4).

Model observer experiments are often limited in size, especially when real data is used. In these cases, it is important to determine the errors in the estimated AUC or SNR. The source of the errors is twofold: variation in test case difficulty (case variability) and variation in estimating the reader performance (reader variability) [Clarkson et al., 2006]. In the terminology of linear model observers, a reader is determined by the template of the model, which we estimate using Eq. 3.10 or Eq. 3.12. Thus, multiple model readers correspond to multiple templates estimated using separate sets of the training images. Typically, we assess model performance within a fully-crossed multiple-reader multiple-case (MRMC) study design which assumes that every reader reads every case. One non-parametric estimate of the variance of AUC in such MRMC study design is the one-shot method defined by Gallas [Gallas, 2006]. The one-shot algorithm gives the estimate of AUC averaged over the readers, hereafter average AUC, and the associated variance. We use these to characterize the performance of the CHO models.

In the course of comparing the CHO models, we also make use of the measure named *statistical efficiency*. Commonly, the relative efficiency η of the current observer characterized by $\mathrm{SNR}_{\mathrm{curr}}$ relative to the reference observer characterized by $\mathrm{SNR}_{\mathrm{ref}}$ is defined as follows

$$\eta = \frac{{\mathrm{SNR}_{\mathrm{curr}}}^2}{{\mathrm{SNR}_{\mathrm{ref}}}^2}. \tag{3.18}$$

The efficiency measure is used within the study to investigate several different parameters of the model observer designs (see Section 3.4.3).

3.3 Methods

We consider three different designs of the channelized Hotelling observer model: (1) a single-slice model (ssCHO), (2) three variants of a multi-slice model (msCHO), one existing and two novel ones, and (3) a volumetric model (vCHO). The models are defined in this section along with the corresponding notation.

The images of interest are 3D data created according to the procedures described in Section 3.2.1. The volume can be thought of as the sequence (stack) of 2D image slices (frames). Further on, the slices correspond to xy-plane (height and width) of the volume and the number of slices determines the z-thickness (in voxels) of the volume, *i.e.*, the number of slices in the sequence determines the thickness of the volume. For simplicity, we assume that the image voxels are isotropic, *i.e.*, they have the same size in each x, y and z direction. For a 3D image $\mathbf{g}$, we denote by N the number of slices in the image and by Q the number of voxels in each slice. Thus, the number of voxels in the image is $M = Q \times N$. The slices of the image are represented as column vectors of Q elements (voxel intensities) – the n-th slice in the sequence is denoted $\mathbf{g}_{(n)}$, $n = 1, ..., N$. To represent the whole image, the slice vectors $\mathbf{g}_{(n)}$ are arranged as an array of column vectors and thus $\mathbf{g} = \left[\mathbf{g}_{(1)}, ..., \mathbf{g}_{(N)}\right]$ is a matrix of the size $Q \times N$.

Remember from Section 3.2.1 (Table 3.1) that our experimental image volumes contain $N = 64$ image slices and $Q = 64^2 = 4096$ voxels per slice, thus the number of voxels in the image is $M = Q \times N = 64^3 = 262{,}144$. These values are of interest later in this section as an illustration of the dimensionality of the calculations involved with different models. Note that actual clinical images are often larger then this. For example, today's flat-panel digital x-ray detectors can produce images of over 2048×2048 pixels in size.

To avoid confusion, we will use 2D-CHO and 3D-CHO to denote the CHO model for 2D and that for 3D images, respectively, without implying any specific CHO design.

3.3.1 Single-slice CHO (ssCHO)

We use the name single-slice CHO to refer to the conventional 2D-CHO [Gallas and Barrett, 2003] when it is run on a single slice in the volume, the central slice of the signal; this is shown in Figure 3.4(a). For example, the signal in our study is centered at the central slice of the image, $\mathbf{g}_{(N/2)} = \mathbf{g}_{(32)}$, and this is the slice of interest for our ssCHO computations. Accordingly, concerning Eq. (3.11), the vector of image data is now $\mathbf{g} = \mathbf{g}_{\mathrm{ssCHO}} = \mathbf{g}_{(N/2)}$ with the number of elements equal to the size of one image slice (in our experiments, $Q = 4096$). Correspondingly, also the size of the channel vectors $\mathbf{u}_p, p = 1, ..., P$ equals Q. The resulting vector $\mathbf{v}$ is the channelized data of the selected image slice, $\mathbf{v}_{\mathrm{ssCHO}} = \mathbf{U}^t \mathbf{g}_{(N/2)}$ of the size P. As an example, the maximum value of P in our ssCHO experiments is 15. Thus, by channelizing

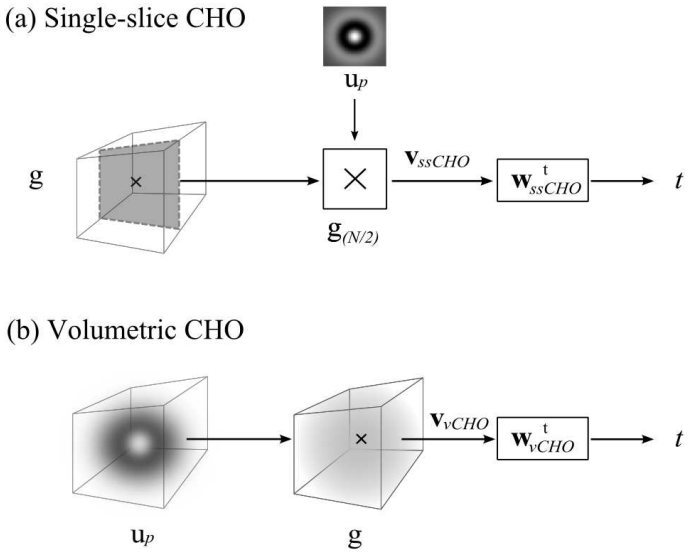


Figure 3.4: (a) Single-slice CHO. The model is constrained to use only the information of one slice in the volume: the slice in which the signal is centered, in our example $\mathbf{g}_{(N/2)}$. First, the $\mathbf{g}_{(N/2)}$ part of an image is channelized using a set of 2D LG channels, $\mathbf{u}_p$, $p = 1, ..., P$ where P is the total number of channels. Then, the vector $\mathbf{v}_{\text{ssCHO}}$ of the channel outputs of the size P is processed by the template $\mathbf{w}_{\text{ssCHO}}$ to estimate the test score t of the ssCHO. (b) Volumetric CHO. The main difference from ssCHO model is that vCHO exploits not only a single slice from the volume but the image volume as a whole, $\mathbf{g} = \left[\mathbf{g}_{(1)}, ..., \mathbf{g}_{(N)}\right]$. Here, the channels match the dimension of the image volume and they are 3D LG functions in a 3D Cartesian space. In any other respect, the vCHO model is the same as ssCHO.

the data, the dimensionality of the data vectors is reduced from $Q = 4096$ down to $P = 15$). Once the channel responses are known, the template of the ssCHO model, $\mathbf{w}_{\text{ssCHO}}$, is estimated using Eq. (3.12).

Since the ssCHO design does not use all image information to perform the detection task, it is expected and proven [Chen et al., 2002, Kim et al., 2004, Wells et al., 2000] to perform not as high as the model designs described next, and for the scope of this work it is used merely as a reference method.

3.3.2 Volumetric CHO (vCHO)

As known from the literature [Barrett and Myers, 2004], the definition of the CHO model is not limited by the dimensionality of the problem as long as the related calculations are manageable. Therefore, a straightforward approach in solving a 3D detection task could remain in the scope of Eq. (3.11), just as it was in the case of ssCHO.

In contrast to the ssCHO which operates only on a single slice $\mathbf{g}_{(N/2)}$ in the image,

the vCHO makes use of the complete image volume, as we depict in Figure 3.4(b). The image vector $\mathbf{g}$ in Eq. (3.11) is obtained by vertically stacking the slice vectors $\mathbf{g}_{(n)}$. Similarly, instead of planar channels of the size Q used in the ssCHO design, we now use volumetric channels represented as column vectors $\mathbf{u}_p$, $p = 1, 2, ..., P$, of the size $M = Q \times N$. By operating on all rather than on a single image slice, the vCHO becomes "aware" of the contrast and correlation between the adjacent image slices which was not the case with the ssCHO.

Specifically, given the fact that the signal in our study is spherically symmetric, we use 3D LG functions which are isotropic in all three dimensions (see Figure 3.3). The 3D LG channels are created in 3D Cartesian space ($\mathbf{r} \in \mathcal{R}^3$), they are of the same size as the image volume, and they are centered on the location of the 3D signal (in our case, the center of the volume).

As with the ssCHO, the template $\mathbf{w}_{\mathrm{vCHO}}$ of the vCHO model is estimated from the channelized data of vCHO using Eq. (3.12). It is important to note that the size of the vector of the channelized data $\mathbf{v}_{\mathrm{vCHO}}$ is the same as that of the $\mathbf{v}_{\mathrm{ssCHO}}$, *i.e.*, equal to the number of channels P. Again, referring to the parameters of our experiments, we note the remarkable reduction in dimensionality brought about by data channelization: the original data vectors $\mathbf{g}$ of $M = 262{,}144$ elements are reduced to a mere $P = 15$ elements of the channelized data vectors $\mathbf{v}_{\mathrm{msCHO}}$. Importantly, note that, in general, we would need at least the number of voxels M in the images to ensure a nonsingular estimate of the data covariance matrix of non-channelized image data (see p.957 of [Barrett and Myers, 2004]). Conveniently, in the domain of channelized image data (channelized models), the number of elements P representing each image is significantly smaller ($P \ll M$) resulting in a markedly relaxed requirements for the number of training images.

3.3.3 Multi-slice CHO (msCHO)

Three different designs of msCHO are considered in this work: type a ($\mathrm{msCHO_a}$), type b ($\mathrm{msCHO_b}$) and type c ($\mathrm{msCHO_c}$). Unlike the ssCHO which exploits information of a single slice only, the multi-slice model design makes use of multiple slices in the image sequence. Similar to ssCHO, yet unlike vCHO, the multi-slice observer makes use of 2D rather than 3D channels to filter the image prior to estimating the linear discriminant (see Figure 3.5).

While this chapter is not directly focused on modeling human observer performance, the design of msCHO model is partly inspired by the postulates about how humans view the volumetric image data sets while using the sequence-browsing image presentation. For example, we may think of a radiologist who is inspecting a multi-slice CT image of the chest. We follow a simplifying assumption of [Chen et al., 2002] that humans interpret the multi-slice image in a two stage process. First, they "pre-process" the image in planar view (xy-plane), slice after slice, and buffer the scores obtained for each slice (hereafter referred to as the *planar scores*). Next, the

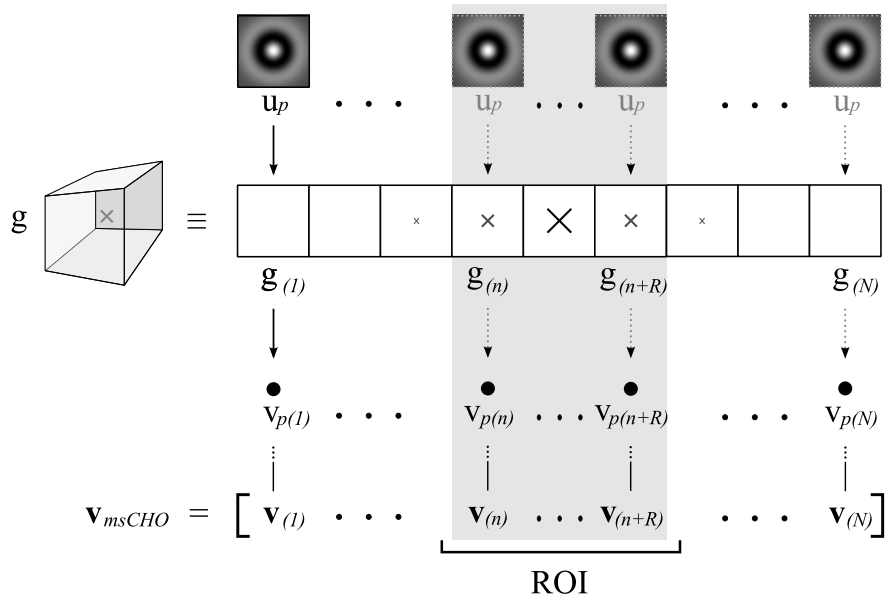


Figure 3.5: Multi-slice CHO. Processing slice data with 2D-LG channels. The multi-slice image $\mathbf{g}$ is represented as an array of slices $\left[\mathbf{g}_{(1)}, ..., \mathbf{g}_{(N)}\right]$, where N is the number of slices in the image. Each slice in the array $\mathbf{g}_{(n)}$ is channelized by the same set of P two-dimensional channels $\mathbf{u}_p$, $p = 1, ..., P$, to get the channel outputs $\mathbf{v}_{(n)} = [v_{1(n)}, ..., v_{P(n)}]$, where $v_{p(n)} = \mathbf{u}_p{}^t \mathbf{g}_{(n)}$. The matrix of the channel outputs for all slices in the image is denoted $\mathbf{v}_{\mathrm{msCHO}} = [\mathbf{v}_{(1)}, \mathbf{v}_{(2)}, \ldots, \mathbf{v}_{(N)}]$. The same procedure applies on both signal-present and signal-absent images. The concept of the region of interest (ROI) is explained in Section 3.3.3.5.

planar scores are processed ("integrated") in the z-direction to result in the sequence test statistic t which is used to make the classification decision: normal or abnormal case. Further on, we refer to these two phases as *pre-processing stage* and *integration stage*, respectively.

3.3.3.1 Pre-processing stage

For all three multi-slice designs, the slice data is first processed with a set of 2D-LG channels, as illustrated in Figure 3.5. Here, each slice in the image, $\mathbf{g}_{(n)}$, $n = 1, ..., N$, is channelized by the set of P two-dimensional channels $\mathbf{u}_p$, $p = 1, ..., P$, to get the channel outputs $\mathbf{v}_{(n)} = [v_{1(n)}, ..., v_{P(n)}]$, where $v_{p(n)} = \mathbf{u}_p{}^t \mathbf{g}_{(n)}$. This resembles ssCHO design only now the channels are applied on each slice in the sequence individually rather than only on the central slice. For simplicity, for each of the N slices

in the sequence, we use exactly the same set of 2D channels.[7] In line with Eq. (3.11), the channelized data of the n-th slice is $\mathbf{v}_{(n)} = \mathbf{U}^t \mathbf{g}_{(n)}$, where $n = 1, ..., N$ and $\mathbf{U}$ is the channel matrix. The matrix of the channel outputs for all slices in the image is denoted $\mathbf{v}_{\mathrm{msCHO}} = \left[\mathbf{v}_{(1)}, \mathbf{v}_{(2)}, \ldots, \mathbf{v}_{(N)}\right]$.

The msCHO models differ in how they use the channelized slice data $\mathbf{v}_{\mathrm{msCHO}}$. In general, two approaches have been taken in handling the $\mathbf{v}_{\mathrm{msCHO}}$; these are illustrated in Figure 3.6. For one approach, applied in the $\mathrm{msCHO_c}$, the output of the pre-processing stage is the channelized data, $\mathbf{v}_{\mathrm{msCHO}}$ (see Figure 3.6(c)). The other approach, applied in $\mathrm{msCHO_a}$ and $\mathrm{msCHO_b}$, extends the pre-processing stage to calculate also a test statistic $t_{(n)}$ for each slice $n = 1, ..., N$. In view of model design, this corresponds to a 2D-CHO which is run on each slice in the sequence to build an array of test statistics for all slices denoted $\mathbf{t}_{\mathrm{planar}} = [t_{(1)}, t_{(2)}, ..., t_{(N)}]$ (see Figure 3.6(a),(b)). Here, $\mathbf{t}_{\mathrm{planar}}$ is considered the vector of planar scores and it is used as input to the subsequent integration stage. The details of the three variants of the model are discussed next.

3.3.3.2 msCHO type a

This model design is illustrated in Figure 3.6(a) and it corresponds to the work of Chen *et al.* [Chen et al., 2002] and Gifford *et al.* [Gifford et al., 2005]. The channelized slice data obtained in the early pre-processing stage is used to estimate 2D-CHO templates at each slice position in the sequence. These templates model the hypothesis that humans examine different slices of the stack with different signal templates in mind. We can write the 2D-CHO template matrix for a given slice position n as

$$\mathbf{w}_{(n)} = {\mathbf{K}_{\mathbf{v}_{(n)}}}^{-1} \Delta\overline{\mathbf{v}_{(n)}}, \quad n = 1, ..., N. \tag{3.19}$$

That is to say, for each slice position n, the template $\mathbf{w}_{(n)}$ is estimated using equally positioned slices of the trainer image sequences. For example, to build a template for the first slice of the tester sequences ($n = 1$) we use only the first slices of the trainer images. As such, there are N different templates $\mathbf{w}_{(n)}$ in total. Note that the dimensionality of the covariance matrices $\mathbf{K}_{\mathbf{v}_{(n)}}$ in Eq. (3.19) is $P \times P$, the same as for the ssCHO model. Therefore, the requirements in terms of the number of training images required for accurate estimation of the data covariance are comparable to those of the ssCHO.

Next, the templates are used to calculate the test statistic for each slice in the planar view. The output data may be summarized in a vector of planar CHO measures, $\mathbf{t}_{\mathrm{planar}} = [t_{(1)}, t_{(2)}, \ldots, t_{(N)}]$, where

$$t_{(n)} = {\mathbf{w}_{(n)}}^t \, \mathbf{v}_{(n)}, \quad n = 1, ..., N. \tag{3.20}$$

[7] Note that using the same channels for all slices in the sequence may be not the most efficient, especially in the cases where the signal properties vary a lot from one slice to another. In those cases, it may be worthwhile tuning the channel parameters for different slice position separately.

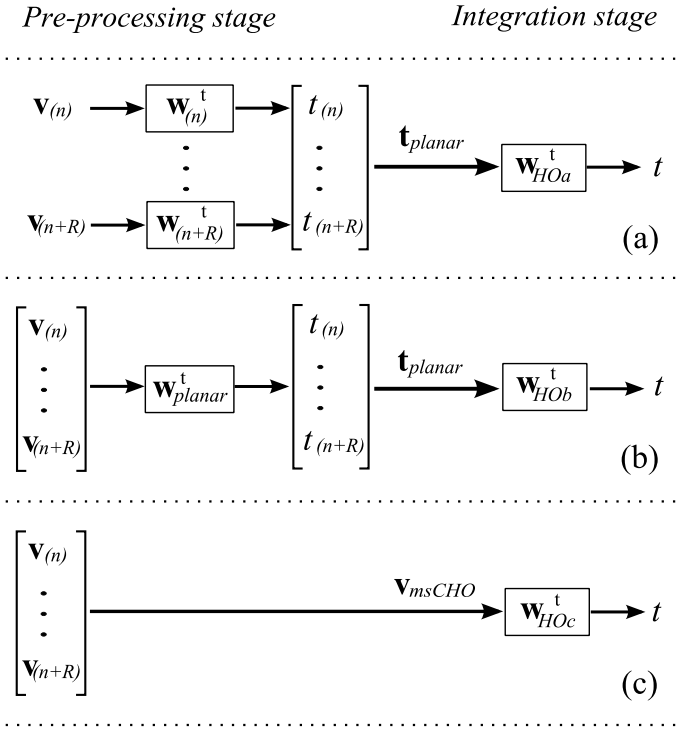


Figure 3.6: Three different designs of the multi-slice CHO: (a) $\mathrm{msCHO_a}$, (b) $\mathrm{msCHO_b}$, (c) $\mathrm{msCHO_c}$. Each observer is applied on the region of interest (ROI), consisting of R consecutive slices where $R \leq N$ and $R = N$ corresponds to the whole image sequence; for details about ROI see Section 3.3.3.5. The three msCHO models process an image in two stages: the pre-processing stage and the integration stage. First, in the pre-processing stage, the channelized slice data $\mathbf{v}_{(n)}, ..., \mathbf{v}_{(n+R)}$ is obtained (see Figure 3.5). In two out of three msCHO designs, (a) and (b), this channelized data is used to calculate the vector of test statistics, $\mathbf{t}_{\mathrm{planar}} = [t_{(n)}, ..., t_{(n+R)}]$, using different templates in Eq. (3.8): (a) a separate 2D template $\mathbf{w}_{(n)}, ..., \mathbf{w}_{(n+R)}$ is used for each of the R processed slices, (b) the same 2D template $\mathbf{w}_{\mathrm{planar}}$ is used for all slices in the ROI (*e.g.* $\mathbf{w}_{\mathrm{planar}} = \mathbf{w}_{(N/2)}$). Thus, the output of the pre-processing stage is $\mathbf{t}_{\mathrm{planar}}$ for (a) and (b), or the vector of channelized slice data $\mathbf{v}_{\mathrm{msCHO}}$ for (c). Subsequently, in the integration stage, these data are processed by the 1D-HO which estimate the final test statistic t for the image sequence.

In the final step, the integration phase, $\mathbf{t}_{\mathrm{planar}}$ is used by the one-dimensional HO to calculate the final scalar statistic of the msCHO model; namely

$$t(\mathbf{t}_{\mathrm{planar}}) = (\mathbf{K}_{\mathrm{planar}}^{-1}\Delta\overline{\mathbf{t}_{\mathrm{planar}}})^t\,\mathbf{t}_{\mathrm{planar}} = \mathbf{w}_{\mathrm{HO}_a}{}^t\,\mathbf{t}_{\mathrm{planar}}. \tag{3.21}$$

It is important to remark that the HO template is also estimated using the trainer data, just as the 2D-CHO templates are. To do this, the 2D-CHO templates from Eq. (3.19) are applied to the trainer images in order to estimate $\mathbf{t}_{\mathrm{planar}}$ vectors for the trainer sequences which are then used to estimate the HO template, $\mathbf{w}_{\mathrm{HO}_a} = \mathbf{K}_{\mathrm{planar}}^{-1}\Delta\overline{\mathbf{t}_{\mathrm{planar}}}$.

3.3.3.3 msCHO type b

In contrast to the $\mathrm{msCHO_a}$ where a different template was used for each of the consecutive slices, we propose the first new model $\mathrm{msCHO_b}$ design using one 2D-CHO template over multiple adjacent slices in the image. This assumes that human observers may be more likely to examine multiple successive slices with a unique signal template in mind. The $\mathrm{msCHO_b}$ model is illustrated in Figure 3.6(b).

The number of consecutive slices to be processed with the same signal template depends on the inter- and intra-slice thickness as well as on the signal properties, especially the signal spread, and the background variability. Ideally, when the slice thickness is small, the background variability is not too high, and the signal characteristics are not changing significantly across slices, a single template could be applied on every slice in the sequence (or in the part of the sequence, see Section 3.3.3.5), independent of the slice position within the sequence [Platiša et al., 2009b]. In view of Eq. (3.20), we shall call this template $\mathbf{w}_{\mathrm{planar}}$, *i.e.*, $\mathbf{w}_{(n)} = \mathbf{w}_{\mathrm{planar}}, n = 1, ..., N$. For simplification, in our study we assume that the aforementioned conditions are approximately satisfied and use a single template for each slice in the image sequence. Specifically, given the location of our 3D signal (centered on the image volume), we estimate the $\mathrm{msCHO_b}$ template by substituting the value of $n = N/2$ in Eq. (3.19), *i.e.*, $\mathbf{w}_{\mathrm{planar}} = \mathbf{K}_{\mathbf{v}_{(N/2)}}{}^{-1}\Delta\overline{\mathbf{v}_{(N/2)}}$. The same as with $\mathrm{msCHO_a}$, the size of the covariance matrix to be estimated from the training data is $P \times P$.

If, on the other hand, the slice thickness would be large (implying little correlation between adjacent slices), the signal would be spread over fewer slices or there would be pronounced disturbances in its isotropy, or the background variability would be large, it might be not correct to apply the same template on all slices in the sequence. Rather, a separate template should be estimated for each subset of "similar" adjacent slices of the testing sequences. Eventually, for the greatest variability of the data, a separate template should be estimated for each slice position in the sequence, hence $\mathrm{msCHO_b}$ would converge to $\mathrm{msCHO_a}$.

Lastly, in the integration phase of $\mathrm{msCHO_b}$ design, the vector of slice test statistics is used by the HO with the template $\mathbf{w}_{\mathrm{HO}_b}$ to infer the image test statistics, $t = t(\mathbf{t}_{\mathrm{planar}})$. This is exactly the same as the integration phase of $\mathrm{msCHO_a}$ design.

3.3.3.4 msCHO type c

Inspired by the work of Young *et al.* [Young et al., 2009], we propose an alternative multi-slice CHO approach and the second novel CHO design in this study $\mathrm{msCHO_c}$. Here, we assume that the humans may be processing the sequence-browsing data by first only filtering the images slice by slice (without making any "per-slice detection" as in $\mathrm{msCHO_a}$ and $\mathrm{msCHO_b}$) and then using all per-slice-filtered data to infer the detection decision for the whole sequence. To model this, the channelized slice data $\mathbf{v}_{\mathrm{msCHO}}$ are fed directly to an HO to integrate into a final observer score for the image, as depicted in Figure 3.6(c).

The $\mathrm{msCHO_c}$ approach is most similar to vCHO in that the correlation between slices are directly incorporated in the model. In the scenario of $\mathrm{msCHO_c}$, the test statistic of the model is

$$t(\mathbf{v}_{\mathrm{msCHO}}) = (\mathbf{K}^{-1}_{\mathrm{msCHO}}\, \Delta\overline{\mathbf{v}_{\mathrm{msCHO}}})^t\, \mathbf{v}_{\mathrm{msCHO}} = \mathbf{w}_{\mathrm{HO}_c}{}^t\, \mathbf{v}_{\mathrm{msCHO}}. \tag{3.22}$$

We notice that the size of the covariance matrix $\mathbf{K}_{\mathrm{msCHO}}$ is determined by the number of slices N in the image and by the number of channels P. As mentioned earlier, P is usually of the order of 10 while N often well exceeds this range. This suggests potential difficulties in estimating the template $\mathbf{w}_{\mathrm{HO}_c}$ in Eq. (3.22) caused by the large dimensionality of the covariance matrix, similar as in [Barrett et al., 2001], especially when the available trainer data set is limited in size. For example, when $N = 64$ and $P = 10$, the number of elements in $\mathbf{K}_{\mathrm{msCHO}}$ is $(N \times P)^2 = 409{,}600$. As suggested in [Barrett and Myers, 2004], there are several alternative approaches to be considered when direct inversion of the covariance matrix is not feasible: iterative computation of the template (using either iterative or regularized methods), Neumann series, or matrix-inversion lemma. For more information, the interested reader is referred to [Barrett and Myers, 2004].

Theoretically, it can be shown that, for the task of detecting a separable signal in a Gaussian background with separable covariance matrix, the three msCHO variants have the same asymptotic SNR (AUC) performance (asymptotic refers to the number of training images which tends to infinity) [Goossens et al., 2012b]. In general, without those specific conditions, $\mathrm{msCHO_c}$ will outperform $\mathrm{msCHO_a}$ (assuming a sufficient number of training images) while $\mathrm{msCHO_a}$ may slightly outperform $\mathrm{msCHO_b}$. We discuss this further in the results section.

3.3.3.5 Region of interest (ROI)

As we have defined them so far, the multi-slice CHO models can use all slices in the image sequence. However, Wells *et al.* [Wells et al., 2000] studied the task of detection of small lesions in thoracic Ga-67 SPECT data and found that the benefit of a multi-slice display comes primarily from the two slices immediately adjacent to the central slice. The authors used a 1 cm diameter sphere to model the signal, where each voxel width was 0.317 cm. We shall henceforth refer to this subset of significant adjacent

slices the region of interest (ROI), where the number of slices in the ROI is denoted R.

The preferred size of the ROI is influenced by the imaging technology (slice thickness and separation), as well as by the statistical properties of image data, including smoothness and symmetry of the signal, the range of its spread over slices, and the variability of the background content. All considered, the value of R shall be chosen to fit the properties of the given data. In the example of the human observer study by [Wells et al., 2000], it was shown that increasing the ROI (in their case $R > 3$) brings less significant improvement in the observer performance.

In our experiments, each of the three msCHO designs illustrated in Figure 3.6 are applied on the ROI of size R for which the channelized slice data is depicted in Figure 3.5. The value of R is varied among the values of 3, 5 and 11 adjacent slices centered around the slice with the peak signal intensity ($n = 32$); the details are discussed in Section 3.5.3.

3.4 Experimental setup

In the following, we first review the parameters of the image data used in the experiments and how the images are grouped (training/testing). Next, we explain the design of the experiments and we end with the summary of the methods used for data analysis.

3.4.1 Sample images

For the experiment setup, the testbed of image ensembles is comprised of four categories: WNB, CNB, LB and CLB, as described in Section 3.2.1. Detailed parameters of all background images are summarized in Table 3.1. The total number of synthesized backgrounds is $22{,}000$ for each WNB and CNB categories, and $14{,}000$ for each LB and CLB categories. The simulation time is significantly longer for LB and CLB compared to LB and CLB. We aim at a data set which is large enough to ensure statistical significance of the results while the computational time and computer power required for both image generation and observer calculations remain within reasonable limits. In each category, half of the set is used as signal-absent images and the remaining half is used to create signal-present images by inserting a 3D Gaussian signal in the center of the background volume (see Figure 3.2). Given the parameters of the background images and aiming at non-trivial detection tasks, the spread of the 3D Gaussian signal is assigned $\sigma_s = 8$ throughout the study. In addition, for CNB data we also consider $\sigma_{s2} = 5$ and $\sigma_{s3} = 3$. For each background category and each considered σ_s, the peak intensity of the signal, a_s, is varied in the range of four different values, selected to approximately fit the criterion of AUC covering the range from 0.6 to 0.9 in equal steps. Due to different parameters of the backgrounds, a_s values differ across four categories, as specified in Table 3.1.

The image data is used as follows. For the WNB and CNB categories, $10{,}000$ pairs (hereafter called *trainer pairs*) of signal-present and signal absent images are used as training data. For the LB and CLB categories, the number of trainer pairs is 6000. In all categories, 1000 image pairs (hereafter called *tester pairs*) are used as test data. Tester data are kept independent from the trainer data.

3.4.2 Study design

We test the performance of five CHO designs: ssCHO, $\mathrm{msCHO_a}$, $\mathrm{msCHO_b}$, $\mathrm{msCHO_c}$, and vCHO, for four image categories: WNB, CNB, LB and CLB. Initially, we run a set of experiments to select the parameters of LG channels. Next, the performance and variance of the CHO models are evaluated in multiple-reader multiple-case (MRMC) studies. For CNB data, we investigate the influence of signal parameters σ_s and a_s. This will allow us to assess the influence of the signal size. In addition, for WNB and CNB images, the IO performance is estimated using Eq. (3.17) and Eq. (3.16). Finally, for the three multi-slice observers, we investigate the influence of ROI size on the model observer performance.

For all considered model observers, the observer templates are estimated using the trainer data. For a given CHO, all template parameters (the covariance matrix $\mathbf{K}$, the mean channelized signal $\Delta\overline{\mathbf{v}}$, the mean planar test statistics $\Delta\overline{\mathbf{t}_{\mathrm{planar}}}$) are estimated using the exact same pairs of signal-absent and signal-present trainer 3D images. In the testing phase, the observer templates are used in estimating the test statistics for each of the tester 3D images. There is no overlap between the trainer and the tester image sets.

As defined in the previous section, 2D channels are required for both ssCHO and msCHO experiments while 3D channels are used by vCHO only. To that end, we explore two basic types of the CHO models: ssCHO to select parameters of the 2D channels and vCHO to select parameters of the 3D channels. Given that the sampled 3D LG channels as used in the study are not exactly orthonormal, we considered also the orthonormalized version of the 3D LG channels. In line with the work of Gallas and Barrett [Gallas and Barrett, 2003], each model is investigated for several values of the channel spread parameter: for $\sigma_{s1} = 8$, $a_u = \{7, 12, 18, 24, 32\}$; for $\sigma_{s2} = 5$, $a_u = \{4, 7, 12, 15, 21\}$; and for $\sigma_{s3} = 3$, $a_u = \{3, 5, 7, 9, 12\}$. For each spread parameter, the number of LG channels is varied in the range of $P = 1, ..., 30$. The experiments are conducted with $N_{\mathrm{tr}} = 2000$ trainer pairs and $N_{\mathrm{ts}} = 1000$ tester pairs, and for the second largest among four considered values of signal magnitude a_s given in Table 3.1. Further in the study, these selected channel parameters are used. Within the same image category, a unique set of 2D LG channels is used for both the ssCHO and msCHO, while the 3D LG channels are used for the vCHO. The exact same set of 2D LG channels are used for all three types of the msCHO and for all slices in the image sequence.

The MRMC studies are characterized by the following parameters: the number of

Table 3.2: Multiple-reader multiple-case (MRMC) study configurations. The total number of each of WNB and CNB images is 11000 image pairs, and the total number of each of LB and CLB images is 7000 image pairs. For all study configurations, the number of tester image pairs is fixed to $N_{\mathrm{ts}} = 1000$. No overlap exists between the trainer images and the tester images.

Background category	Number of trainer image pairs, N_{tr}	Number of readers, N_{rd}
WNB, CNB	$\{50, 100, 200, 500, 1000, 2000\}$	5
	$\{5000\}$	2
LB, CLB	$\{50, 100, 200, 500, 1000\}$	5
	$\{2000\}$	3

trainer image pairs (N_{tr}), the number of tester image pairs (N_{ts}) and the number of readers (N_{rd}).The exact values of these parameters are given in Table 3.2. A range of different values of N_{tr}, while N_{rd} and N_{ts} are kept fixed, will allow the influence of the size of trainer data set to be evaluated. Each of the specified MRMC configurations is repeated for every signal spread value, σ_s, and related range of four signal magnitudes, a_s, all as specified in Table 3.1.

3.4.3 Figures of merit

The performance measures used in the study include: AUC, the estimate of AUC variance, SNR, and the model efficiency η; these are all defined in Section 3.2.3. For each MRMC configuration, we first estimate AUC and then use it in Eq. (3.16) to calculate the SNR. To evaluate the variability associated with the results, we use the one-shot variance analysis [Gallas, 2006]. Eventually, in analyzing the influence of particular parameters of the CHO designs on their performances, we focus on CNB category of the data and use Eq. (4.11) to estimate efficiency, η, of the observers.

Three different types of the model observer efficiency are considered: efficiency of the CHO relative to the IO, η_{CHO}, efficiency of the CHO trained with fewer trainer pairs relative to its performance for the largest considered number of trainer pairs, $\eta_{\mathrm{N_{tr}}}$, and efficiency of the ssCHO model relative to the vCHO model, $\eta_{\mathrm{ss,v}}$. In view of Eq. (4.11), the actual SNR values used in place of $\mathrm{SNR_{curr}}$ and $\mathrm{SNR_{ref}}$ for each different type of the efficiency are specified in Table 3.3.

3.5 Results and discussion

To facilitate the interpretation of the results, we first refer to the IO performance in 2D versus 3D detection tasks and analyze the influence of image parameters and level of task difficulty on the performance gap between the two. We then proceed to elaborate

Table 3.3: Terms of Eq. (4.11) for three different types of model observer efficiency, $\eta = {\mathrm{SNR}_{\mathrm{curr}}}^2/{\mathrm{SNR}_{\mathrm{ref}}}^2$

Type of efficiency	$\mathrm{SNR}_{\mathrm{curr}}$	$\mathrm{SNR}_{\mathrm{ref}}$
η_{CHO}	SNR of a given CHO model	SNR of the IO
$\eta_{\mathrm{N_{tr}}}$	SNR of the CHO trained with N_{tr} image pairs, $N_{\mathrm{tr}} < 5000$ (see Table 3.2)	SNR of the CHO trained with maximum considered number of trainer pairs, $N_{\mathrm{tr}} = 5000$
$\eta_{\mathrm{ss,v}}$	SNR of the ssCHO model	SNR of the vCHO model

on the selection of the channel parameters used in the study and continue to present a detailed comparative analysis of the five CHO models described in Section 3.3. Finally, we point to the major differences among these models and think about their potential applications in the future.

3.5.1 Difficulty of the detection task: 2D versus 3D

Before we get into the analysis of the CHO model performances, it is worthwhile looking at the performance of the IO for the 2D versus the 3D problem, respectively, 2D-IO versus 3D-IO. In Figure 3.7, we show these results for the two image categories in the study for which the data is Gaussian: WNB (top graph) and CNB (bottom graph). As stated in Section 3.2.2.1, when the image data are Gaussian the ideal linear observer, the HO, equals the IO. We first calculate the SNR of the IO using Eq. (3.17) and then use this to calculate the AUC of the IO by inverting Eq. (3.16).

Looking at Figure 3.7, we notice that the 3D-IO outperforms the 2D-IO for both the WNB and CNB. Such results confirm our intuition about the gain in the observer performance from using the information from more than a single slice in the detection process. This is in line with the fact that 3D observer, unlike the 2D one, exploits also the information about signal contrast in z-direction and about spatial correlation structure between the slices of the data which yields more accurate estimates of the signal $\mathbf{s}$ and the covariance matrix $\mathbf{K}$ in Eq. (3.17).

Moreover, the difference between 2D-IO and 3D-IO performance is much more significant in the case of WNB compared to the CNB images. This may be explained by different levels of difficulty of the detection tasks in the two categories of image data. Namely, going from 2D to 3D adds more information on the signal which results in the 3D-IO outperforming the 2D-IO. However, when there is correlation in the backgrounds, such as in CNB, 3D also adds more complexity to the background which makes the detection task more difficult and diminishes the positive impact of the additional signal information. Having together the benefit that comes from extra signal

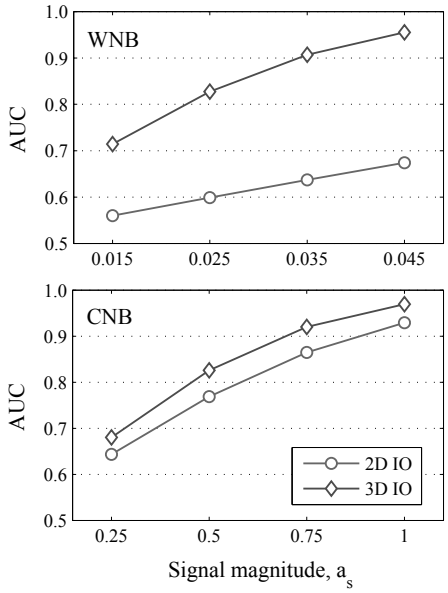


Figure 3.7: The IO performance for two Gaussian image categories: (Top) WNB and (Bottom) CNB. The two curves in each graph correspond to a two-dimensional and a three-dimensional problem, respectively, 2D IO and 3D IO. The 2D images are of size $M = 64^2$ with a 2D Gaussian signal inserted in the center of the image, while the 3D images are of size $M = 64^3$ with 3D spherically symmetric Gaussian signal inserted in the center of the volume. For both 2D and 3D Gaussian signal, the value of the signal spread parameter is $\sigma_{s1} = 8$. Further details about image parameters are given in Table 3.1. The AUC values are obtained using Eq. (3.17) to calculate SNR and then Eq. (3.16) to calculate the AUC of the IO.

information and the detriment that comes from increased background complexity, the performance difference between 2D and 3D is narrower when the backgrounds have correlation, *i.e.*, the detection task is more difficult. Admittedly, the statistical complexity of real clinical images surpasses those of the data in our study, especially the WNB. Even so, the remarks about the possible factors of larger or smaller benefits of 3D over 2D remain relevant also for clinical applications.

Clearly, we expect the performance trends among ssCHO and vCHO model designs to be similar to those of the IO, both in terms of the 2D versus 3D approach and uniform versus inhomogeneous image contents (backgrounds).

Last, we note that the difficulty of the detection task, either 2D or 3D, depends not only on the correlation of the background data but also on other parameters of image objects. For example, in our study the signal is of Gaussian shape with the spread $\sigma_s = 8$ for both WNB and CNB while the spread of the CNB Gaussian kernel is determined by $\sigma_b = 8$. Given that the size and shape of the signal are the same as

those of the filter kernel for the noise, the CNB detection task may be described as difficult. In contrast, the task is relatively easy for the WNB. When, for example, we would decrease the signal size (*e.g.* $\sigma_s = 5$) the difficulty of the WNB task would increase while the difficulty of the CNB task would decrease. Indeed, by looking at the parameters of CNB data in Table 3.1, we observe that the decrease in the signal spreads, $\sigma_{s1} > \sigma_{s2} > \sigma_{s3}$, is followed by the decrease in the signal amplitudes, $s_{s1} > s_{s2} > s_{s3}$, while the background structure is fixed, $\sigma_b = \sigma_{s1}$. This decreasing trend in the level of the signal, while preserving the value of the AUC, confirms the decrease in the difficulty of the detection task [Burgess, 1999a].

3.5.2 Exploring channel parameters

On the way to evaluate the CHO models, we first run a series of experiments for each of the four image categories aiming to select the parameters of 2D and 3D LG channels such that they capture as much information as possible for the purpose of signal detection. The results of this investigation for $\sigma_s = 8$ are depicted in Figure 3.8. Here, the graphs in the left and right column depict results for the ssCHO and vCHO, respectively, while the rows represent the image categories: WNB, CNB, LB and CLB, from top to bottom. Each curve in a graph corresponds to a different a_u. The solid lines labeled "ideal observer" show the AUC performance of the IO calculated using Eq. (3.17) and Eq. (3.16): 2D-IO for 2D image data and 3D-IO for 3D image data. Further in the text, IO is used to refer to 3D-IO unless otherwise indicated. The number of trainer images used in these experiments is $N_{\mathrm{tr}} = 2000$ which allows a meaningful estimate of the involved data covariance matrices. For each image category, the selected number of 2D LG channels is denoted P_{2D} and the selected number of 3D LG channels is denoted P_{3D}.

We observe in all plots that the curves nicely converge to an asymptote as the number of channels increases from 1 to 30. Looking in more detail, for both ssCHO (2D LG) and vCHO (3D LG), the curves for narrower channels reach a higher performance with fewer channels but then approach the asymptote more slowly. For wider channels, on the other hand, performance improves more gradually as the number of channels increases but does not have the long approach to the asymptote. Further on, we notice that in most categories ssCHO converges faster than vCHO, reaching to the asymptote with a fewer channels hence there $P_{3D} > P_{2D}$. We found these results independent of the orthonormality of the 3D LG channels. In line with the task difficulty discussion at the beginning of this section, the distance between the CHO asymptote and the IO score is more pronounced in the case of ssCHO compared to vCHO where the linear model nearly approaches the IO.

Aiming at the best and stable performance of the CHO with a reasonable number of channels and given the plots in Figure 3.8, we select the channel parameters which are used further in the study. For all four image categories and related signal size, the selected parameters of the LG channels are listed in Table 3.4, these particular

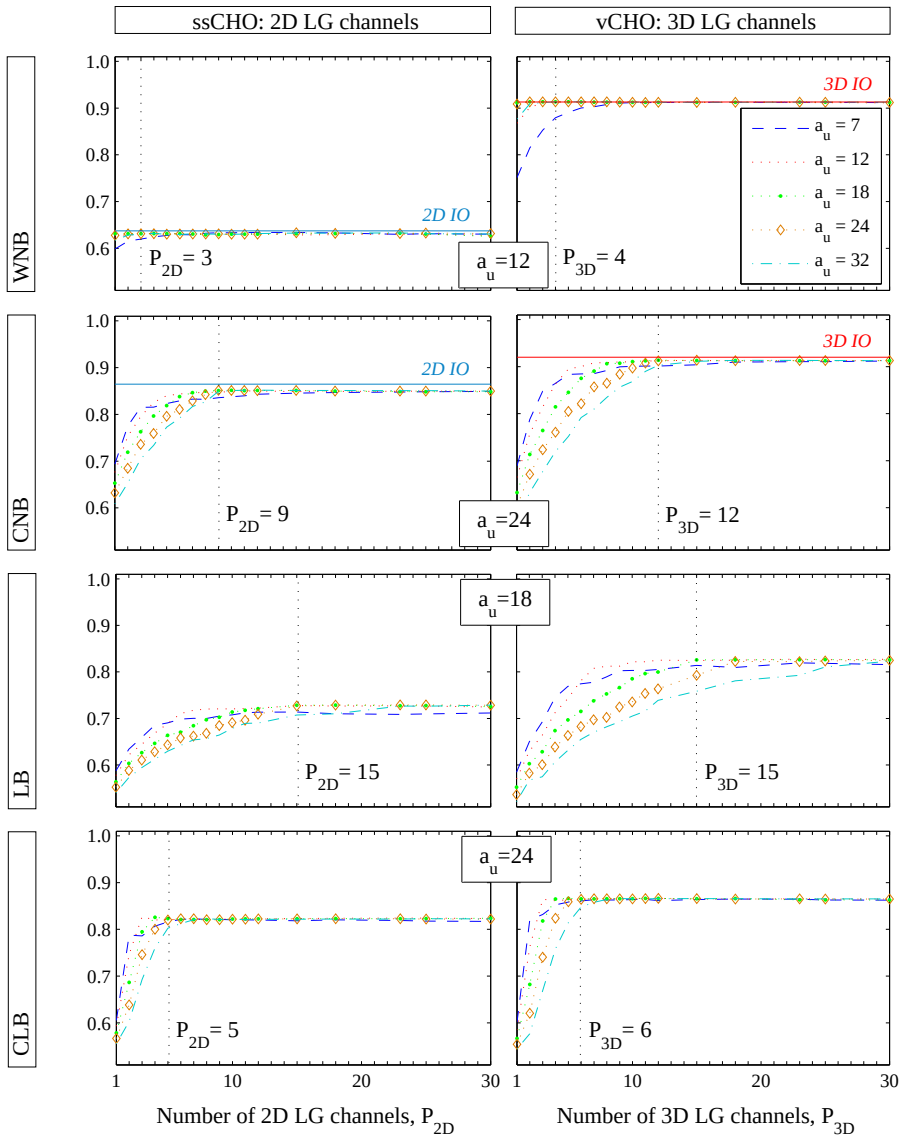


Figure 3.8: Plots of the estimated AUC as a function of a number of channels, $P = 1, \ldots, 30$. Right: AUC for the ssCHO design using 2D LG channels applied on the central slice of the image sequence. Left: AUC for the vCHO design using 3D LG channels. For both model designs, a set of different spread parameters is considered, $a_u = \{7, 12, 18, 24, 32\}$. Top to bottom: The results for WNB ($a_s = 0.035$), CNB ($a_{s1} = 0.75$), LB ($a_s = 12$), and CLB ($a_s = 12$). The plots are obtained for $N_{\text{tr}} = 2000$ trainer image pairs and $N_{\text{ts}} = 1000$ test image pairs. Selected channel parameters are listed in Table 3.4: channel spread parameter a_u, and number of channels P_{2D} for ssCHO, and P_{3D} for vCHO.

Table 3.4: Parameters of the Laguerre-Gauss (LG) channels. For each background category and the related signal size, parameters of the LG channels are determined: the size of the channels, a_u, the number of 2D LG channels, P_{2D}, and the number of 3D LG channels, P_{3D}. The parameters of 2D and 3D LG channels are selected in the experiments with ssCHO and vCHO models, respectively. The models are investigated in the space of five families of LG channels defined by the value of the channel spread parameter, $a_u = \{7, 12, 18, 24, 32\}$. For each family, the number of LG channels is varied in the range of $P = 1, ..., 30$. The experiments are conducted with $N_{\mathrm{tr}} = 2000$ trainer pairs and $N_{\mathrm{ts}} = 1000$ tester pairs, and for the second largest among four considered values of signal magnitude a_s given in Table 3.1. The results of these experiments are illustrated in Fig. 3.8.

Background category	Signal size	a_u	P_{2D}	P_{3D}
WNB	$\sigma_s = 8$	12	3	4
CNB	$\sigma_{s1} = 8$	24	9	12
	$\sigma_{s2} = 5$	21	11	12
	$\sigma_{s3} = 3$	12	12	12
LB	$\sigma_s = 8$	18	15	15
CLB	$\sigma_s = 8$	24	5	6

values of a_u, P_{2D} and P_{3D} are used in the remainder of the study. With respect to the category, the narrowest and fewest channels are used in case of WNB while wider and more of those are used for other image categories. Again, the tendency conforms with the difficulty of the detection tasks. Thus, for example, $a_u = 12$, $P_{2D} = 3$, $P_{3D} = 4$ for WNB while for more complex CNB these values increase to $a_u = 24$, $P_{2D} = 9$, $P_{3D} = 12$.

3.5.3 Comparing CHO performances

The performance results for the five CHO model designs are summarized in Figure 3.9 for all four categories of image background and for the specific data parameters defined in Table 3.1. The signal size is the same in all images, $\sigma_s = 8$.

The results correspond to the study design of $N_{\mathrm{tr}} = 2000$ trainer pairs and $N_{\mathrm{ts}} = 1000$ tester pairs, and $N_{Rd} = 5$ readers (templates) for WNB and CNB backgrounds, or $N_{\mathrm{rd}} = 3$ readers for LB and CLB data, all according to Table 3.2. The details about the MRMC study configurations can be found in Section 3.4.2.

For the msCHO models the size of ROI is $R = 11$ with approximately 65% of the signal energy included in the decision process. Here, the energy of the signal is calculated as $E(\mathbf{s}) = \sum_{m=1}^{M} s_m^2$ where $\mathbf{s}$ is defined by Eq. (3.4) and M is the number

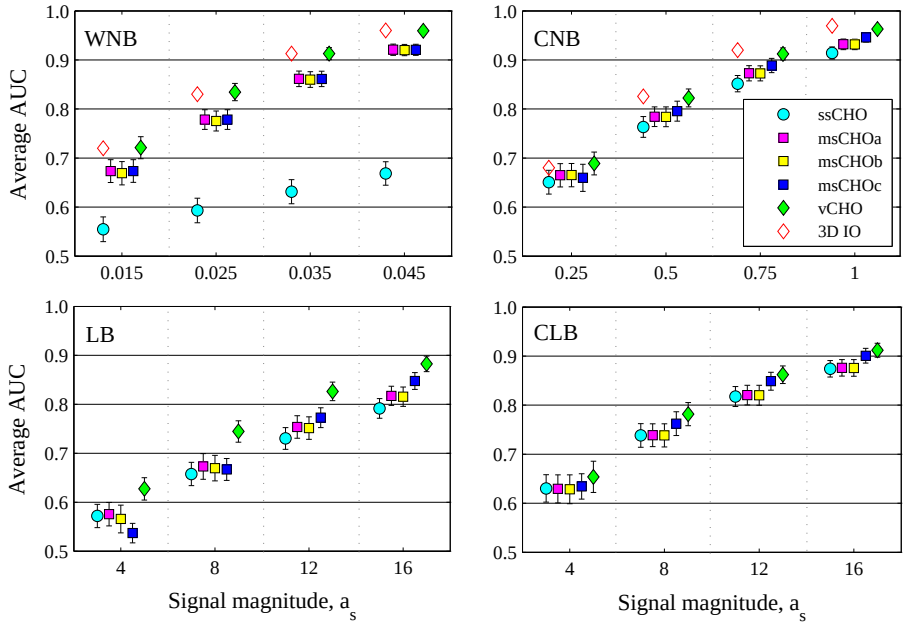


Figure 3.9: Average AUC for the five model observer designs: ssCHO run on the central slice in the image; $\mathrm{msCHO_a}$, $\mathrm{msCHO_b}$ and $\mathrm{msCHO_c}$ each applied on the region of interest comprised of $R = 11$ adjacent slices centered on the central signal slice; vCHO applied on the whole image volume. Each graph corresponds to one of the four background categories (left to right, top to bottom): WNB, CNB, LB and CLB. The value of signal spread parameter is $\sigma_{s1} = 8$ and the related signal magnitudes a_s correspond to those defined in Table 3.1. Number of trainer image pairs per reader $N_{\mathrm{tr}} = 2000$, and number of tester pairs $N_{\mathrm{ts}} = 1000$. Number of readers N_{rd} corresponds to the applicable study configurations from Table 3.2. Error bars are ± 2 standard deviations estimated by the one-shot method [Gallas, 2006].

of voxels in the signal image. The size of ROI is selected such that the $\mathrm{msCHO_c}$ covariance matrix can be well estimated (see discussion in Section 3.3.3.4).

In each experiment, the AUC is averaged over the total number of readers. The error bars are ± 2 standard deviations estimated by the one-shot method [Gallas, 2006]. For the purpose of this analysis, and in view of the remarks from Section 3.4.1 concerning the selection of the signal magnitudes, we shall avoid directly comparing absolute values of the AUC for different image categories. Instead, we compare relative trends in AUC values of the different CHO variants only within the same image category.

In all four data categories, vCHO clearly outperforms the other models. Among multi-slice designs which are ranked next, the $\mathrm{msCHO_c}$ which infers the classification decision directly from the channelized slice data $\mathbf{v}_{\mathrm{msCHO}}$, outperforms the other two

which use $\mathbf{v}_{\mathrm{msCHO}}$ to build the slice test statistics prior to estimating the final image statistic. On the lower side, expectedly, is the ssCHO design. For all five models, the error bars slightly decrease as the magnitude of the signal grows.

Across the four image categories, the most striking difference between the model performances is observed for the WNB images where ssCHO performs significantly worse than the other four models. As explained earlier, the reason for this remarkable benefit of using information from multiple slices in the process of signal detection stems from the low difficulty of the detection task. Even more, given the uniform structure of the white noise WNB and the relatively "large" spread of the signal used in our study ($\sigma_s = 8$), the detection task gets relatively "easy" as the observer gets access to all three-dimensions of the image. The least amount of disagreement between the model performances is observed for CLB images which use the most complex backgrounds in the study.

In further analysis and discussion, we focus on CNB data and explore the influence of specific parameters: signal size, signal magnitude and size of trainer data set.

The results of the MRMC studies for CNB images when the signal size is $\sigma_{s2} = 5$ and $\sigma_{s3} = 3$ are presented, respectively, in the top and bottom graphs of Figure 3.10. For the msCHO, the size of ROI is the same as in Figure 3.9, $R = 11$. The approximate percent of signal energy included in the decision process is now 88% for σ_{s2} and 99% for σ_{s3}. Overall, in Figure 3.10 we observe similar tendencies in CHO model performances as those in Figure 3.9. Only now the absolute difference between performances of the different models is more pronounced.

Let us first look at the ssCHO versus the vCHO. We refer to the results for CNB with σ_{s1} from Figure 3.9 and those for CNB with σ_{s2} and σ_{s3} from Figure 3.10. For example, let us examine the experiment setups when the AUC of vCHO is in the range of 0.9 (the second largest a_s for a given σ_s). By comparing these, we observe that the absolute differences in performance of vCHO and ssCHO for a given σ_{si}, denoted $\mathrm{D}|\sigma_{si}$, $i = 1, 2, 3$, are ordered as follows: $(\mathrm{D}|\sigma_{s1} \approx 0.07) < (\mathrm{D}|\sigma_{s2} \approx 0.2) < (\mathrm{D}|\sigma_{s3} \approx 0.22)$. Earlier in this section, we established that the difficulty of the detection task in three CNB image setups, each with the kernel size of $\sigma_b = 8$, is highest for $\sigma_{s1} = 8$, lower for $\sigma_{s2} = 5$ and lowest for $\sigma_{s3} = 3$. Here again, the ordering of performance differences nicely agrees with the earlier discussion that the benefit of vCHO over ssCHO is most significant when the task difficulty is low ($\mathrm{D}|\sigma_{s3} \approx 0.22$), and it gets smaller for more difficult tasks ($\mathrm{D}|\sigma_{s1} \approx 0.07$). Similar trends appear with respect to the difference between the ssCHO and msCHO. There also, the difference in performance is largest when the task is of lowest difficulty (σ_{s3}).

Another interesting aspect to these results is the influence of ROI size, the number of slices used with the msCHO models. Even for σ_{s3} when 99% of the signal energy is included in the ROI, there is a difference between the msCHO models and vCHO. This may indicate that the msCHO still has insufficient information on background statistics. The extent of the vCHO is not limited to the ROI size. It is possible that the msCHO performance can be increased by choosing more slices but still fewer

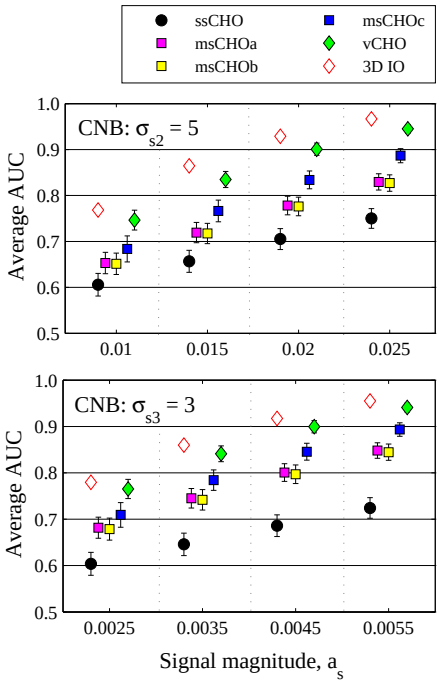


Figure 3.10: For the CNB image category, average AUC of the five CHO model designs and the ideal 3D IO when the value of signal spread parameter is (top) $\sigma_{s2} = 3$ and (bottom) $\sigma_{s3} = 5$. The related signal magnitudes a_s correspond to those defined in Table 3.1. The three msCHO models are applied on the region of interest comprised of $R = 11$ adjacent slices centered on the central signal slice. The number of trainer image pairs per reader is $N_{\mathrm{tr}} = 2000$, the number of tester pairs is $N_{\mathrm{ts}} = 1000$, and the number of readers is $N_{\mathrm{rd}} = 5$. Error bars are ± 2 standard deviations estimated by the one-shot method [Gallas, 2006].

than the whole volume. On the other hand, especially with $\mathrm{msCHO_a}$ and $\mathrm{msCHO_b}$, involving more slices that have little or no signal in them might only add unnecessary noise. Of course, the specific choice of the ROI size should represent the best compromise between the aforementioned considerations. Moreover, it would depend on the type of image data and its background statistics, and for $\mathrm{msCHO_c}$, on the number of training images available to adequately estimate the covariance matrix. Eventually, we remind that in our experiments the same 2D LG channels are used for each slice of a given image sequence which may not be optimal. The influence of the ROI size will be discussed later in the section, yet detailed analysis in this respect requires future research.

We continue with comparing CHO performances to the IO over a range of CNB image parameters σ_s and a_s. By doing this, we aim at evaluating the range of disparity

Table 3.5: Efficiency of CHO models applied on CNB images with different spread of the signal: efficiency of CHO model relative to the IO performance, η_{CHO}; efficiency of ssCHO relative to the vCHO performance, $\eta_{\mathrm{ss,v}}$. Three different values of signal spread parameter are considered: $\sigma_{s1} = 8$, $\sigma_{s2} = 5$ and $\sigma_{s3} = 3$. For each σ_s the exact same backgrounds are used and their lump spread parameter is $\sigma_b = 8$. For multi-slice CHO models (msCHO), the efficiency for the ROI size of $R = 11$ are given. The values of η_{CHO} and $\eta_{\mathrm{ss,v}}$ are calculated using the formula in Eq. (4.11) and as explained in Section 3.4.3. The calculations are done for MRMC configuration with the number of trainer image pairs $N_{\mathrm{tr}} = 5000$. See text for the discussion.

σ_s	a_s	ssCHO η_{CHO} [%]	$\mathrm{msCHO_a}$ η_{CHO} [%]	$\mathrm{msCHO_b}$ η_{CHO} [%]	$\mathrm{msCHO_c}$ η_{CHO} [%]	vCHO η_{CHO} [%]	$\eta_{\mathrm{ss,v}}$ [%]
8	0.25	69	85	86	91	>100	62
	0.5	59	71	71	82	98	60
	0.75	55	66	66	77	93	59
	1	53	63	63	75	91	59
5	0.01	13	28	27	35	82	16
	0.015	13	27	27	36	78	17
	0.02	14	27	26	37	77	18
	0.025	14	26	26	37	76	18
3	0.0025	12	36	35	46	88	13
	0.0035	12	36	36	47	86	14
	0.0045	12	36	35	47	85	14
	0.0055	12	36	35	47	85	15

among different CHO models. It is not the explicit focus of this study to select a CHO model which approximates the IO; the IO model is used as a point of reference. The statistical efficiency of the five CHO model observers relative to the IO, η_{CHO}, is calculated using Eq. (4.11) with the corresponding SNR values from Table 3.3. The results are summarized in Table 3.5.

In general, the definition of SNR from Eq. (3.17) suggests a linear increase of SNR with the increasing signal magnitude, a_s. Thus, for each σ_{s1}, σ_{s2} and σ_{s3}, we expect the efficiency η_{CHO} as defined in Eq. (4.11), to be constant with respect to a_s. Indeed, for CNB setups with $\sigma_{s2} = 5$ or $\sigma_{s3} = 3$, and given the results in Table 3.5, the efficiencies of the CHO models relative to the IO are approximately constant with the considered values of a_s. However, the efficiencies observed with $\sigma_{s1} = 8$ do not meet the expectations. Even, with very low $a_s = 0.25$, it happens that $\eta_{\mathrm{vCHO}} > 100\%$ which, in theory, is not possible. Such unstable behavior of the efficiency η_{vCHO} in the case of σ_{s1} could be attributed to the effect of training the CHO models (see later discussion of $\eta_{\mathrm{N_{tr}|a_s}}$ and Table 3.6).

In comparing the η_{CHO} across three values of σ_s, we notice that the benefit of the vCHO over the other models is more significant for smaller σ_s, *i.e.*, for lower diffi-

Table 3.6: Efficiency of five CHO models for different levels of the signal a_s while the number of trainer images increase: $\eta_{\mathrm{N_{tr|a_s}}}$. For CNB images, the efficiency of CHO models: ssCHO, $\mathrm{msCHO_a}$, $\mathrm{msCHO_b}$, $\mathrm{msCHO_c}$, and vCHO, trained with fewer image pairs relative to their performance for the largest considered number of trainer images, $\eta_{\mathrm{N_{tr|a_s}}}$, are calculated using the formula in Eq. (4.11) and as explained in Section 3.4.3. For three msCHO models, the efficiency for the ROI size of $R = 11$ are given. The rows labeled "$\mathrm{SNR}_{N_{\mathrm{tr}}=5000}$" give the SNR values for $N_{\mathrm{tr}} = 5000$.

a_s	0.25	0.5	0.75	1	0.25	0.5	0.75	1	0.25	0.5	0.75	1
$\mathrm{SNR}_{N_{\mathrm{tr}}=5000}$	0.55	1.02	1.48	1.94	0.70	1.31	1.92	2.53				
	$\eta_{\mathrm{N_{tr\|a_s}}}$ [%] for ssCHO				$\eta_{\mathrm{N_{tr\|a_s}}}$ [%] for vCHO							
$N_{\mathrm{tr}} = 50$	41	74	84	88	55	82	87	88				
100	56	81	89	92	61	84	90	93				
200	69	91	95	97	73	92	96	97				
500	89	96	98	99	90	97	98	99				
1000	96	99	99	100	94	99	99	100				
2000	99	100	100	100	99	100	100	100				
$\mathrm{SNR}_{N_{\mathrm{tr}}=5000}$	0.61	1.12	1.62	2.11	0.61	1.12	1.62	2.11	0.63	1.20	1.75	2.30
	$\eta_{\mathrm{N_{tr\|a_s}}}$ [%] for $\mathrm{msCHO_a}$				$\eta_{\mathrm{N_{tr\|a_s}}}$ [%] for $\mathrm{msCHO_b}$				$\eta_{\mathrm{N_{tr\|a_s}}}$ [%] for $\mathrm{msCHO_c}$			
$N_{\mathrm{tr}} = 50$	13	51	69	75	23	58	72	78	0	0	3	5
100	34	71	81	85	37	70	82	87	11	26	36	42
200	44	83	91	93	54	86	93	95	16	43	60	68
500	78	94	96	97	81	94	97	98	41	70	80	85
1000	88	97	98	99	89	97	98	99	61	85	92	94
2000	97	99	99	100	97	99	99	100	85	95	97	98

culty of the detection tasks. This is confirmed with the calculations of the efficiency of ssCHO relative to the vCHO, $\eta_{\mathrm{ss,v}}$ aimed to illustrate the difference in observer efficiency caused by the restricted amount of information used by the ssCHO compared to the vCHO model design. These are also included in Table 3.5. The value of $\eta_{\mathrm{ss,v}}$ varies significantly from approximately 60% for $\sigma_{s1} = 8$ to approximately 17% for $\sigma_{s2} = 5$ or 14% for $\sigma_{s3} = 3$.

Knowing that the limited size of image ensembles is often encountered with sets of real clinical data, it is important to evaluate the influence of the number of trainer pairs N_{tr} on the efficiency of the CHO models. To do that, we calculate CHO efficiencies relative to the scores obtained with the largest considered trainer data set. Eventually, the greater the efficiency of the model for a smaller value of N_{tr}, the less the CHO depends on the number of available trainer images and the better it suits experiments with a limited number of images. Specifically, we use $\eta_{\mathrm{N_{tr}}}$ to investigate the influence of the size of trainer data set in a twofold manner: with reference to the signal spread parameter, $\eta_{\mathrm{N_{tr}|a_s}}$ – for all ssCHO, msCHO and vCHO models; and with reference to the size of ROI, $\eta_{\mathrm{N_{tr}|R}}$ – for the three msCHO models.

The results of $\eta_{\mathrm{N_{tr}|a_s}}$ calculations for all five CHO models and different levels of the signal a_s are given in Table 3.6. Here, all experiment parameters correspond to the results in Figure 3.9. The rows labeled "SNR$_{N_{\mathrm{tr}}=5000}$" give the SNR values for $N_{\mathrm{tr}} = 5000$. These are included to indicate the absolute range of the observer performance for different signal levels, a_s. We notice that the efficiency $\eta_{\mathrm{N_{tr}|a_s}}$ for ssCHO and vCHO are greater than those of the msCHO models. This difference is more noticeable for lower levels of the signal ($a_s = 0.25$) and it gets less significant for higher signal levels ($a_s = 1$). Also, for each observer model, the values of $\eta_{\mathrm{N_{tr}|a_s}}$ significantly increase with increasing N_{tr} for lower signal levels, and this variability is greatly reduced for higher signal levels. Hence, the influence of the size of trainer data set is less significant when the observer performance is higher. Given the parameter values in our study, the CHO models are most sensitive to the size of trainer data set when $a_s = 0.25$ where $SNR_{N_{\mathrm{tr}}=5000}$ is below 1, and they are least sensitive to the value of N_{tr} when $a_s = 1$ where $SNR_{N_{\mathrm{tr}}=5000}$ is in the range of 2 or greater. This is in line with the conclusions from [Fukunaga and Hayes, 1989] who discussed the effect of finite sample size on training a classifier showing that the bias is a function of the performance level.

For the msCHO models and CNB images with $\sigma_{s1} = 8$ and $a_s = 0.75$, we vary the size of ROI among 3, 5 and 11 adjacent slices and for each of them we calculate $\eta_{\mathrm{N_{tr}|R}}$. Additionally, the $\mathrm{msCHO_a}$ and $\mathrm{msCHO_b}$ models are applied on all slices in the image, $R = 64$. These results are presented in Table 3.7 where columns denote the size of ROI and rows indicate the number of trainers. In case of $\mathrm{msCHO_c}$, the covariance matrix of channelized slice data, $\mathbf{K}_{\mathrm{msCHO}}$ in Eq. (3.22) is of the greatest dimension, $(R \times P)^2$ compared to R^2 of the other two models. When the number of slices in ROI increase to $R = 64$ and given that $P_{2D} = 9$, the size of our data set ($N_{\mathrm{trMAX=5000}}$) is insufficient to properly estimate $\mathbf{K}_{\mathrm{msCHO}}$. Thus, for $\mathrm{msCHO_c}$ the

analysis is restricted to the lower three values of R.

In Table 3.7, we first observe the row labeled "SNR$_{N_{\mathrm{tr}}=5000}$" where the SNR values for $N_{\mathrm{tr}} = 5000$ are presented. For $R = 3$ and $R = 5$, SNR$_{N_{\mathrm{tr}}=5000} = 1.49$ for either $\mathrm{msCHO_a}$ or $\mathrm{msCHO_b}$. Also, from Table 3.6 we read that the ssCHO performance for the same image parameters ($a_s = 0.75$) is SNR$_{N_{\mathrm{tr}}=5000} = 1.48$.

These two SNR values are nearly the same suggesting little benefit for $\mathrm{msCHO_a}$ or $\mathrm{msCHO_b}$ from incorporating the additional 3 or 5 slices adjacent to the slice on which the ssCHO runs. In the case of $\mathrm{msCHO_c}$, the contribution of the first few slices around the signal is slightly greater yet notably less compared to those of $R = 11$. All in all, from the results presented in Table 3.7, it is clear that the major contribution in msCHO performance comes from the next few slices, mainly from the ROI of 11 consecutive slices centered around the central slice of the 3D signal. Further growing the ROI might be considered to fine tune R for a given data. To that end, we note that $\mathrm{msCHO_c}$ is able to reach SNR$_{N_{\mathrm{tr}}=5000} = 1.75$ already with $R = 11$ while the other two models need all $R = 64$ slices to approach this level of the performance.

Eventually, we evaluate the overall influence of the number of trainer pairs on the model performances. As noted before, for all three msCHO designs the efficiency degrades as the size of ROI grows. However, this is more pronounced for fewer trainer pairs and it gradually disappears as N_{tr} grows. Looking back at Table 3.6 and together with Table 3.7, $\mathrm{msCHO_a}$ and $\mathrm{msCHO_b}$ are less sensitive to N_{tr} than $\mathrm{msCHO_c}$. Even further, the $\mathrm{msCHO_b}$ compared to $\mathrm{msCHO_a}$ appears slightly more robust to the changes of ROI size especially when N_{tr} is in the lower range. To illustrate this, when $a_s = 0.75$ and $R = 11$, the $\mathrm{msCHO_b}$ achieves $\eta_{\mathrm{N_{tr}}} > 70\%$ with $N_{\mathrm{tr}} = 50$ but then progresses to $\eta_{\mathrm{N_{tr}}} > 90\%$ already with $N_{\mathrm{tr}} = 200$. The $\mathrm{msCHO_a}$ is a few percent lower while $\mathrm{msCHO_c}$ is able to reach $\eta_{\mathrm{N_{tr}}} > 90\%$ only with $N_{\mathrm{tr}} = 1000$ trainer image pairs, which is in line with the earlier remarks about dimensionality restrictions of the latter model. The least affected by the limited number of trainer images are vCHO and ssCHO models, reaching over 80% of efficiency with as few as $N_{\mathrm{tr}} = 50$.

3.5.4 Some practical considerations

In conclusion of this section, we refer to the potential applications of volumetric versus multi-slice versus single-slice observer designs in the actual 3D detection tasks.

Based on the results of our study, vCHO approaches the IO scores most closely. Therefore, it comes forward as the preferred model for optimization of the system to maximize detection of the 3D signal. In contrast to vCHO, ssCHO performs the worst among all five CHO models in terms of actual performance measures. Still, it follows the trends of the other models, and it is the simplest and fastest to apply. Consequently, it might be considered for preliminary experiments in 3D detection tasks, especially when the initial parameter space is large and shall be downsized prior to further in-depth analysis.

Another important aspect to consider when selecting the preferred CHO design

Table 3.7: Efficiency of msCHO models for different size of ROI while the number of trainer images increase: $\eta_{\mathrm{N_{tr|R}}}$. For CNB images, the efficiency of msCHO models: $\mathrm{msCHO_a}$, $\mathrm{msCHO_b}$, and $\mathrm{msCHO_c}$, trained with fewer image pairs relative to their performance for the largest considered number of trainer images, $\eta_{\mathrm{N_{tr|R}}}$, are calculated using the formula in Eq. (4.11) and as explained in Section 3.4.3. In particular, the efficiency for the signal magnitude of $a_s = 0.75$, each for four different ROI size, $R = \{3, 5, 11, 64\}$, are presented. Here, $R = 64$ implies that the CHO is applied to all slices in the image. The row labeled "$\mathrm{SNR}_{N_{\mathrm{tr}}=5000}$" gives the SNR values for $N_{\mathrm{tr}} = 5000$.

R	3	5	11	64	3	5	11	64	3	5	11	64
$\mathrm{SNR}_{N_{\mathrm{tr}}=5000}$	1.49	1.49	1.62	1.73	1.49	1.49	1.62	1.73	1.60	1.65	1.75	-
	$\eta_{\mathrm{N_{tr\|R}}}$ [%] for $\mathrm{msCHO_a}$				$\eta_{\mathrm{N_{tr\|R}}}$ [%] for $\mathrm{msCHO_b}$				$\eta_{\mathrm{N_{tr\|R}}}$ [%] for $\mathrm{msCHO_c}$			
$N_{\mathrm{tr}} = 50$	84	83	69	19	82	81	72	25	55	36	3	-
100	88	87	81	49	87	86	82	52	77	68	36	-
200	95	94	91	68	95	94	93	73	88	80	60	-
500	97	97	96	88	97	97	97	87	94	89	80	-
1000	99	99	98	94	99	99	98	94	98	96	92	-
2000	100	100	99	98	100	100	99	98	99	99	97	-

are characteristics of the signal. Throughout this study, the signal is a spherically symmetric, isotropic 3D Gaussian function. In practice, however, this would most often not be the case. Certainly, as the signal gets more anisotropic the choice of LG channels as we use them in the study might not be adequate and alternative channels shall be considered. One known alternative are the steerable channels proposed by [Goossens et al., 2010] and used by [Zhang et al., 2012] to detect multiple sclerosis lesions in brain MRI data. Another example from literature are the channels created by [Michielsen et al., 2013] to mimic the non-isotropic point spread function of the DBT images; they were used for detecting micro-calcifications in DBT images.

Even more so, when the signal anisotropy is in the z-direction, perhaps even due to the increased slice thickness, it might be desirable to reconsider not only the channel selection but also the preferred model design for a given application. It may well be that the vCHO design which seems to be the most efficient design in the case of an isotropic 3D signal compares differently to the msCHO designs when the signal characteristics are changed. In particular, when the signal is spread over a very limited number of slices only or its isotropy in z-direction is noticeably distorted, we favor further investigating the msCHO models (see also discussion in Section 3.3.3.3 and Section 3.3.3.5).

Last but not least, given the possible applications of the model observers from this study, we refer to the CHO model designs from the perspective of mimicking humans. While anthropomorphic models as such are outside the scope of this work, we refer to some of their basic considerations to stimulate the discussion. As proposed by Myers and Barrett [Myers and Barrett, 1987], the property of frequency selective channels which are known to exist in the HVS is used to model the process of signal detection in the two-dimensional environment. This mechanism certainly extends to three-dimensional problems. For video imaging applications, for example, it has been modeled with a three-dimensional filter bank which is separable in spatial and temporal frequency components [Van den Branden Lambrecht, 1996]. However, current literature does not tell us how exactly the HVS is channelizing the data when viewing it in sequence browsing mode where the speed of browsing is not predetermined and the forward-backward looping is allowed. Conveniently, the sequence-browsing viewing scenario itself resembles the technique of msCHO signal detection. Henceforth, it might be worthwhile to further explore the msCHO model designs to better understand their relation to the human performance. This direction of research may also benefit from the findings of [Wolfe, 2013] who studied the question of when does a human leave one scene (here slice) for the next one.

Even more challenging is the design of anthropomorphic models which operate on real clinical images. Inevitably, a number of factors have to be considered here, ranging from the anatomical properties of the signal as well as of the background, through the parameters of the underlying imaging technology (inter- and intra-slice reconstructed thickness), the speed of browsing through the sequence, and the limitations of the medium of image presentation such as the temporal effects in slow

medical displays. Again, considerations about robustness of the model designs to the number of trainer images may play an important role in applications dealing with real clinical images where a limited number of samples are available. In addition, an important aspect of modeling human observers is the issue of channel selection. Finally, most of the existing models assume the SKE condition which clearly does not correspond to clinical practice.Especially, the issue of signal location uncertainty is now becoming more researched [Zhang et al., 2012, Gifford, 2013, Lau et al., 2013, Leng et al., 2013, Popescu and Myers, 2013]. As [He and Park, 2013] point out, the questions related to the mechanisms of the signal search (next to the signal detection) may find some answers in the vision research. An example is the work of [Pelli, 1985] which defines the uncertainty model of visual detection by combining two assumptions about visual detection: probability summation (the signal may be detected for many reasons and any of these is sufficient for the successful detection) and decision-variable (effects of the subjective criterion). Undoubtedly, in-depth further investigations are necessary before the preferred design of the anthropomorphic 3D model can be proposed.

3.6 Conclusion

This work studied candidate model observers for the task of signal detection in a 3D problem. We considered three models previously used in the literature (ssCHO, $\mathrm{msCHO_a}$, and vCHO) and two novel models ($\mathrm{msCHO_b}$ and $\mathrm{msCHO_c}$). In that sense, we have presented the theoretical background for the selected models and conducted an experimental comparative analysis of those for a range of statistically different images. Where applicable, the models were compared to the IO known to set the theoretical boundary for the signal detection performance.

Throughout our experiments, the signals were known exactly (spherical isotropic Gaussian blobs centered on the image volume) and the backgrounds were known statistically (the statistical complexity varying from uniform Gaussian white noise, through Gaussian correlated noise, to non-Gaussian lumpy and clustered-lumpy backgrounds). For all image categories, the CHO using volumetric channels was outperforming the other model designs. Even more, when the data statistics were Gaussian, the vCHO closely approached the scores of the IO. Accordingly, the vCHO seems a good candidate for a practical implementation of a an efficient 3D model observer, a model which can approximate the ideal linear observer performance. Next ranked were the multi-slice observers, where the novel proposed $\mathrm{msCHO_c}$ performed the best, followed by the $\mathrm{msCHO_a}$ and the $\mathrm{msCHO_b}$. This ordering of the msCHO performance is in line with the theoretical findings from [Goossens et al., 2012b]. Moreover, for conditions corresponding to our Gaussian image data, it can be shown theoretically that the three msCHO variants have the same asymptotic detection performance. Along the assumptions about human visual system which motivated the design of three msCHO models, this concept of CHO appears as a candidate for an-

thropomorphic model design. Finally, on the low end of the detection performance scale was the ssCHO, as expected. Importantly, the disparity between the models became less pronounced as the difficulty of the task (determined by the parameters of the image objects) increased.

Further on, we found that the major benefit of multi-slice versus single-slice observer comes from the several adjacent slices centered around the signal referred to as ROI, rather than from all slices in the sequence. This agrees with the conclusions in [Wells et al., 2000]. The exact size of the ROI is subject to the properties of a particular data set (slice thickness, signal spread, background statistics, etc.) and shall be determined on a case-by-case basis. Among msCHO designs, the new $\mathrm{msCHO_b}$ seemed least affected by the number of training samples, assuming the size of ROI was appropriately selected. By its design, in particular the relatively large size of the covariance matrix, the $\mathrm{msCHO_c}$ model was most sensitive to the size of training ensemble and thus most susceptible to the dimensionality problem.

The msCHO models were further explored in the context of two studies for task-based quality evaluation of medical display systems; those were reported in Chapter 4 of this thesis. Moreover, in that same chapter, we reported about the human observer study which explored the detection performance in single-slice versus multi-slice (sequence-browsing) image presentation. That study was aimed to serve as a preliminary guide for modifications towards msCHO models which could better predict human performance.

The contributions reported in this chapter have resulted in one journal paper [Platiša et al., 2011e] and three international conference publications, two as the first author [Platiša et al., 2009b, Platiša et al., 2009a] and another as a co-author [Goossens et al., 2012b].

4

Observer studies for medical displays

One key factor of medical image viewing is the *quality of the medical display* system. When developing a new medical display, or approving it for the market, or making a decision on which clinical display to buy for the hospital, it is most important to assess the clinical value of the display, *i.e.*, how well it can serve the *clinical task* of interest. Typically, this task-based quality evaluation is done through human or model observer studies.

In this chapter we present the related studies conducted during this dissertation. Specifically, there are four studies using the model observers which have been defined in Chapter 3 (for three of these studies relevant human data exist and we also refer to those) and one study with human observers. The experiments consider either two-dimensional or three-dimensional images, or both. Also, both clinical and synthetic images are explored.

4.1 Introduction

What is very different about medical displays in comparison to the standard commercial electronic displays is the purpose for which they are used. Unlike the commercial devices, such as television sets and computer monitors, medical display systems are meant to serve the medical processes (*e.g.* establishing a clinical diagnosis, performing a surgery on a patient, or doing a routine clinical review). Because their target applications are different, also the expectations of the two categories of the displays are very different. In the case of commercial displays, the priority is often the overall degree of "excellence" of an image [Engeldrum, 2004]. We refer to this kind of image quality (IQ) as the technical IQ (TechIQ), or *beauty* of the image. Given the large number of manufacturers and devices in the market, the problem of image quality assessment (IQA) for this type of the display systems has been widely studied and there are even standards for the related procedures. As an example, we mention recommen-

dation BT.500-13 by ITU-R entitled "Methodology for the subjective assessment of the quality of television pictures" [ITU-R, 2012].

In contrast, in the case of medical displays, it is essential to evaluate the *utility* (usefulness) of the displayed images rather then their beauty. While it is definitely desirable that the images look appealing to the clinicians (the end user of the displays), it is in fact the utility of the displayed images for the specific clinical task that determines the quality of the medical display systems. Therefore, when assessing the quality of a medical display, it is essential that we evaluate the level of clinical performance for the given display. Commonly, clinical task performance is assessed by means of the so-called *observer studies* in which the observers (either humans or models) perform the task of interest.

Ideally, in the high technology world of today, we would like to have medical imaging processes fully automated and that includes also IQA of the medical displays (*e.g.* in the sense of IQA for real-time control of the parameters affecting IQ). This implies clear preference for model- over human observers. To their advantage, models save financial costs of the IQA,[1] but also tremendously shorten the IQA process and allow the faster introduction of new technology into the clinical setting. For illustration, a typical human observer study with medical specialists may take some months, often even longer. The complete process includes image data preparation, training (getting the observers acquainted with the images, the task, and the exact test procedure), actual image readings, and statistical data analysis. The same process applies for the model observer studies only there training (*e.g.* CHO template estimation, see explanation in Section 3.2.3) and testing take considerably less time. Several factors contribute to the time efficiency of the models, including, obviously, the markedly smaller processing time per image but also the benefit of not depending on observers' availability (can be a major factor of delay, especially with expert human observers).

As we suggested in Section 3.1, model observers (*e.g.* the 2D-CHO model) have been used previously to guide imaging system development and optimization. Even more, the model observer studies reported here demonstrate the value of the models in assessing the quality of medical displays, not only for planar image viewing but also for the case of 3D sequence-browsing. Nonetheless, the models are still not mature enough to completely take over the human studies.

Indeed, in practice, it is well-accepted to use model observer studies throughout the life-cycle of a medical display: starting from the design and prototyping, through optimization, all the way to the preclinical validation. The step where human observer studies exclusively replace the models is the final clinical validation of a new product. The reason is simply the unacceptably high cost of a potential error of allowing to the market a technology (device, algorithm, imaging agent) which is not well-tested and

[1] At the SPIE Medical Imaging conference 2011, during the workshop entitled "Device Evaluation: Perspectives from Inside and Outside the FDA", a "six-figure" number (referring to US dollars) was suggested as an illustration of a typical cost of a human observer study for clinical validation of a new medical device, a display or another.

safe for the patients – the consequences could be disastrous. Though considerably improved over the years, the model observers need to be rigorously evaluated (and possibly refined) before they could be allowed to completely replace humans. Again, considering the seriousness of the risk involved, this has to be done gradually, for different tasks and different image modalities.

Finally, we discuss the "absolute" versus "relative" performance of a model observer. When a model observer study is used for the purpose of a display evaluation, it is often not necessary that it predicts an absolute level of the performance of humans. Rather, this type of study typically aims at comparing a new product with an already existing and well-tested one, old product. Therefore, it is of interest to quantify the difference between the performance of two products. For instance, suppose that a model observer study would estimate the AUC values of the new and the old product to be 0.78 and 0.7, respectively. This would suggest that the new product is of higher utility than the old one. While it is possible (or even likely) that the corresponding AUC values of humans would not be the same (*e.g.* these could measure 0.75 and 0.71, respectively), the literature suggests that the overall ranking of the two products would most often be the same and so it is possible to use models to predict the "relative" performance of the two products. In our example, we would find that the new product is of higher clinical utility than the old product, both for the humans and for the models; despite the difference in the "absolute" performance values. Consequently, even though the models often outperform humans in absolute terms (the exact AUC values), they can be successfully used to assess the utility of the new products.

The work presented in this chapter has been performed within the framework of the "Medical Virtual Imaging Chain" (MEVIC) project financially supported by iMinds which involved collaboration with Barco N.V., Belgium. In the course of different experimental studies, we closely collaborated with Dr. Aldo Badano and Dr. Brandon D. Gallas (U.S. Food and Drug Administration, USA), Dr. Cédric Marchessoux and Dr. Tom Kimpe (Barco N.V., Belgium), Prof. Karel Deblaere M.D. (Department of Neuroradiology, Ghent University Hospital, Belgium), Prof. Bart Goossens, Dr. Ewout Vansteenkiste, and Asli Kumcu (Department of Telecommunications and Information Processing, Ghent University, Belgium).

We start this chapter with an overview of the basic concepts around the observer studies, model or human, and some basic remarks about the software platform which we use for our experiments. Next, we present detailed reports on five observer studies for task-based IQA of medical displays: a study with two-dimensional (2D) x-ray chest images, three studies focusing on the effects of the slow response time of the medical liquid crystal display (LCD) systems, and a human observer study evaluating the impact of the image data properties on signal detectability both in single-slice and in multi-slice image viewing. At the end, we draw some conclusions of the work.

4.1.1 The basic concepts of observer studies

In the context of task-based image quality assessment (IQA), the concept of an *observer study* (or a *reader study*) is used to refer to experiments with either human observers (*e.g.* medical doctors) or (non-ideal) model observers (*e.g.* the variants of the channelized Hotelling observer (CHO) model discussed in Chapter 3). Obviously, the key elements of an observer study are the following:

1. the *image data* which is being evaluated;
2. the *observers* (humans or models) who "observe" (operate on, interpret, read, inspect) the image data in order to perform the task of interest.

Commonly, an observer study is conducted in a *multiple-reader multiple-case* (MRMC) design where multiple observers (readers) act on multiple images (cases) [Gallas et al., 2009]. Specifically, when all readers read all cases, the study is referred to as a *fully-crossed* MRMC study design. All observer studies in this dissertation are of this type. A detailed list of the parameters which characterize an MRMC study is provided in Table 4.1. As an example for further explanations, we consider the case of our study reported in Section 4.2: an observer study to assess the image quality (IQ) of the two-dimensional (2D) chest radiography images where the task of interest is detection of lung nodules.

We discuss the observers first. In the case of a human observer study, the observers (subjects) can be either experts for a given task (experienced radiologists) or novices (radiologists in training) or even observers who are "naive" (unfamiliar) to the task (non-medical experts or students). In the scope of this dissertation, the effects of human observer expertise and experience on the IQA are explored in Chapter 2. In the area of model observers, the level of human skills can be as a result of training the model (classifier). If three models have been trained on three different sets of image data – the set with a large, a medium, and a small number of images, then these models could be seen as an expert, a novice, and a non-expert (naive) observer, respectively. Thus, another concept of interest is the level of *expertise* of a human observer, or correspondingly, the *training* of a model observer. The concept of training is discussed later. Note for now that the aforementioned paradigm of an MRMC study design refers to already "trained" observers.

Another important aspect related to the observers is the number of different "entities" in the study – persons in a human study, or different non-overlapping sets of the trainer images in a model observer study. In the MRMC terms, this corresponds to the number of readers, N_{rd} (see Table 4.1). For human observers, it is intuitively clear that individuals are likely to differ in average task performance, more so when the images are of non-trivial diagnostic cases (*e.g.* subtle lung lesions or lung lesions obscured by rib structure). Similarly, performance of model observers may vary (more or less) depending on the specifics of training (the variability in content and the number of images used to estimate the CHO templates, *i.e.*, to build the classifiers). This is ex-

actly the reason for which observer studies are designed as multi-reader experiments measuring the average rather than the individual performance.

Clearly, multiple readers require also the analysis of variability and the influence of the *number of observers* on the estimated task performance. Typically, the number of readers for a human observer study (assuming a given number of image test cases) is chosen based on the sample size estimates from the *pilot study* analysis targeted at certain statistical power[2] [Hillis and Berbaum, 2005, Hillis et al., 2005, Hillis, 2007]. For example, in the study reported in Section 4.6.3 we chose the number of readers/cases targeting at 80% statistical power. In general, the smaller the variability in the measured performance of observers (persons or models), the fewer entities are needed in order to achieve a statistically significant assessment of IQ, and the other way around, the larger the variability, the more observers are needed. At the same time, a "large" number of observers is not easy to achieve, neither with humans nor with model observers. On the one hand, experts (such as experienced chest radiologists) are expensive and often not readily available. On the other hand, as we will discuss shortly, a large number of model observers (CHO templates) often requires a large number of training images which are often not available, especially not in the case of experiments with real clinical image data.

Now, we turn to the image data related concepts in the observer studies (see rows two three in Table 4.1). Irrespective of the type of observers, human or model, the task is performed for a set of images referred to as *tester data*. In MRMC terms, these are the cases which the readers read. Obviously, a meaningful estimate of reader performance requires multiple test cases, hence the term multiple-case experiments. Overall, an estimate based on a large number of cases is more reliable than an estimate based on a few cases. Accordingly, one more parameter to consider when planning an observer study is the *number of tester images*. Specifically, when the number of signal-absent images is equal to the number of signal-present images (the prevalence is 50%), we commonly express the size of the tester image set in number of *pairs* of tester images denoted N_{ts}. Here, a pair is comprised of one signal-present and one signal-absent image sample.

Lastly, we get back to the aforementioned concept of the training which is indispensable for model observer studies but also important for human studies. In particular, in the case of CHO models from Chapter 3, the model observer first has to be trained, *i.e.*, the template of the model has to be estimated based on the training image data. Remember from Chapter 3 that multiple MRMC readers in the context of a model observer study differ only in their CHO templates: each of these is estimated from a different (sub)set of the training images. As with the tester images, we describe the training data in terms of the number of *trainer image pairs*. The difference is in that here we refer to the number of image pairs *per reader* denoted

[2]For more details, the interested reader is referred to the "Sample Size Estimation Overview" by the Medical Image Perception Laboratory of the University of Iowa, `http://perception.radiology.uiowa.edu/SampleSize/tabid/182/Default.aspx`.

N_{tr}. Eventually, the total number of image pairs in an MRMC study with N_{rd} readers equals $(N_{\mathrm{tr}} \times N_{\mathrm{rd}}) + N_{\mathrm{ts}}$. Depending on the exact values, this may add up to a quite large number.[3] However, as already mentioned, large numbers of images are rarely available and compromises have to be made in the study design.

While not the point of investigation for this dissertation, a training process is commonly also involved in human observer studies. Before human observers perform the task for the test images, they would be presented with a number of images (usually not overlapping with the image instances used for the actual performance computations). Nevertheless, the purpose of such training is usually (but not always) for human observers to familiarize themselves with the images and the actual perceptual task, rather than to actually take them to a certain level of "expertise" for the task. Accordingly, the number of trainer images in human studies is relatively small compared to that of model observer studies. For instance, in our study with diagnostic veterinary pathologists reported in Chapter 2 we used as few as 5 images to train our observers. On the other hand, if the task is "simple" and requires no extensive knowledge and special medical competence, it can be possible for naive observers to be trained for the task in a reasonably short period of time and with a manageable number of images. An example could be the study reported in Section 4.6 of this chapter where the task was to detect a Gaussian signal buried in the images of correlated Gaussian noise. The number of trainer images in that study was of the order of 100. In the literature, there are also studies with notably more trainer images, *e.g.*, [Gallas, 2001] trained his observer with over 7000 images (viewed in pairs, one signal-present and the other signal-absent). Clearly, as with model observers, more extensive training of humans suggests less variability in their performance. On the other hand, it requires more of the observer's time and effort, which is a well-known limiting factor of the studies with humans in general.

Overall, observer study design is determined by a range of parameters, and it is very important to evaluate and understand their effects on the estimated level of observers' task performance. As noted earlier, explicit parameters of MRMC design are the number of readers and the number of test cases but not the level of experience or the number of trainer images. An important implication of this is that statistical analysis of the MRMC performance (*e.g.* the one-shot method used in our experiments [Gallas, 2006]) informs about the effects of reader and test image variability but not about the effects of training image variability (training effects). Rather, those have to be examined separately.

In the case of models, this is done by repeating the same MRMC experiment for different values of the parameter N_{tr} (see for examples the experiments in Chapter 3 and in Section 4.2). In the case of human observer studies, there are two training components: the knowledge, skills and experience of an individual observer acquired

[3] As an illustration, we refer to the model observer experiments from Chapter 3. There, the number of readers $N_{\mathrm{rd}} = 5$, the number of tester image pairs $N_{\mathrm{ts}} = 1000$ and the number of trainer image pairs per reader $N_{\mathrm{tr}} = 2000$ making a total of $5 \times 2000 + 1000 = 11{,}000$ pairs of images for the whole study.

Table 4.1: Key parameters of observer studies

Domain of interest	Parameter name	Symbol
MRMC study	Number of readers	N_{rd}
(human or model observers)	Number of tester image pairs	N_{ts}
Model observer study	Number of trainer image pairs	N_{tr}
Human observer study	Expertise and experience	-

at any point before the study (which could be directly related to the task or not, and which in itself is not easy to quantify) and those acquired during the (comparatively very short) training process of the specific study. Analysis of training effects in the context of human observer studies is out of the scope of this thesis.

To summarize, the following requirements apply to the images used in an observer study (human or model):

1. all images (tester and trainer) should be of known class membership (the ground truth is known): signal-absent or signal-present;

2. they should be representative of the problem: the image objects (background, signal, noise) as well as the proportions between the number of signal-absent and signal-present cases should be typical of the population under study;

3. ideally, there must be no overlap between the set of the trainer and the set of the tester images, nor between the subsets of the trainer images used for different CHO readers (multiple CHO templates) in order to ensure independent readers.

4.1.2 Simulation platform

All model observer experiments described in this chapter are implemented and performed within the framework of the *Medical Virtual Imaging Chain* (MEVIC) software platform developed by the company Barco N.V., Belgium [Marchessoux et al., 2008c]. MEVIC is implemented in the C++ programming language and it allows simulations of a complete medical imaging chain. Specifically, the framework includes three major groups of modules:

1. imaging modules, *i.e.*, models of image objects and tools for image collection and ground truth storage;

2. display modules, *i.e.*, models of medical displays (controlled by a range of technology variables) together with visualization software and a number of common image processing algorithms;

3. modules for IQA, *i.e.*, the CHO models from Chapter 3 and the related analysis tools, such as the "one-shot" algorithm for analysis of variance of the area under the ROC curve (AUC) [Gallas, 2006] (the common figure of merit for model observers).[4]

With regards to the development and optimization of medical imaging devices, the aim of the MEVIC platform is twofold: (1) to provide an alternative for physical prototyping of a new product (by simulating the relevant hardware modules) and (2) to avoid involving medical experts in preclinical product validation (by using the model observers instead). In this way, the process of parameter optimization in the product development stage would be significantly more efficient in terms of both time reduction and cost saving.

4.2 Medical displays for chest radiography

We advocate in Chapter 3 that the parameters of a model observer study need to be carefully chosen for the particular image data at hand (the background, the signal, and the noise) in order to avoid misleading conclusions, under- or overestimation of the observer performance, *i.e.*, of the utility of the images. As outlined in Section 4.1.1, the key parameters of a model observer study include the number of readers, the number of trainer images per reader, the number of tester images, and in the case of channel-based models, also the parameters of the channels (*e.g.* the spread and the number of Laguerre-Gaussian (LG) polynomials). In Chapter 3, we explore in detail how different parameters affect the performance of the models on the synthetic images in which the variations exist but those are known. Here, we augment those investigations by looking into the effect of different parameters of a model observer study with real clinical images for which we have little or no knowledge about the underlying data statistics.

In particular, this study addresses the case of medical displays for chest radiography applications. The task of interest is the detection of subtle lung nodules in chest radiographs. We investigate the performance of the ssCHO model (conventional 2D-CHO model) with LG channels with regards to the parameters of the channels, the number of the trainer images, and the number of readers. Eventually, we reflect on a related human observer study as a point of reference in assessing the applicability of the ssCHO model in optimization and evaluation of medical imaging displays.

4.2.1 Study rationale

As noted in Chapter 3, early model observer studies often focused on simulated image data of low complexity. The main reasons for using the simplistic data models

[4]The implementation of the one-shot algorithm is our slightly modified version of the code developed by Matthew A. Kupinski.[5] The one-shot module of MEVIC has been verified against the results of Brandon D. Gallas, the author of the method [Gallas, 2006].

rather than the real clinical images include the following: (1) full knowledge of the data statistics, which allows analytical data analysis, (2) possibility of controlling the parameters of image objects (*e.g.* correlations in the background, presence of noise, signal parameters), (3) known image class membership (signal-present versus signal-absent) and exact location(s) of the abnormalities (if present), and (4) fast and easy access to an arbitrarily large number of images (because the images are computer generated). Undoubtedly, the studies with synthetic images are most useful, not only in the process of observer model development but also in revealing the basic mechanisms of human visual perception; some relevant examples of such studies include [Burgess et al., 1982, Myers et al., 1985, Rolland and Barrett, 1992, Eckstein et al., 1997, Burgess, 1999b, Abbey and Barrett, 2001].

Nevertheless, when it comes to practical applications, it is often of interest to study the observer (human or model) performance with real clinical images. More recently, there has been a growing body of literature reporting observer studies with clinical image data. For instance, [Shidahara et al., 2006] used real clinical SPECT images and reported good agreement between performance of a CHO and human in the task of detecting Alzheimer's dementia. In addition, numerical observer studies have been performed for the data sets acquired by inserting simulated lesions in clinically acquired image data of healthy subjects [Samei et al., 1999, Samei et al., 2003, Chawla et al., 2007]. On the other hand, recent research results in the area of system optimization applications indicate significant potential for the model observers to replace humans. For example, [Chen and Barrett, 2005] used model observers to aid the lens design for a digital mammography system. Also, [Chawla et al., 2008] rely on model observers to build an optimization scheme for multi-projection breast imaging.

4.2.2 Experimental goal

It is well-known from the literature that the performance of observers on signal detection tasks in medical x-ray images is limited by both *quantum noise* (the variations caused by the finite number of x-ray photons producing the image) and *object variability* ("anatomic noise", the variations formed by the projection of anatomic structures, such as ribs, vessels, organ tissue) [Samei et al., 1999, Barrett, 1990]. The primary objective of our study is to evaluate the performance of the ssCHO model in the task of subtle lung nodule detection in real clinical chest x-ray backgrounds (chest radiographs). Specifically, we consider the ssCHO model with LG channels and explore the influence of the following parameters on the model performance: the spread of the LG channels, the number of the LG channels, the number of the trainer images, and the number of readers. Moreover, we are interested in evaluating the potential for the single-slice CHO model (ssCHO) model to be used in optimization and evaluation of the medical displays for chest radiography.

(a) Normal chest radiograph 2048 x 2048 pixels

(b) Test image: signal-absent 508 x 508 pixels

(c) Test image: signal-present 508 x 508 pixels

Figure 4.1: Randomly selected examples of the image data used for the study: (a) a normal clinical image of 2048×2048 pixels in size taken from the data set of [Shiraishi et al., 2000] and used to extract (crop) the background images of the size of 420×420 pixels, (b) a signal absent image used in the study, (c) a signal present image used in the study. The white markers in (c) indicate the line chosen to depict the profile of pixel gray level values shown in Figure 4.2.

4.2.3 Study design and methodology

In this section, we describe the main aspects of the experimental study design: (1) the process of creating the images, (2) the way the images are used in the model observer experiments, and (3) the methods of data analysis.

4.2.3.1 Image data

The test images were prepared in several steps. First, background images were extracted from the healthy (lesion-free) clinical radiographs of the data set from [Shiraishi et al., 2000]. Caution has been taken that the rib edges were positioned off the image center (this will prevent the masking effect – interference with the signal inserted later to form the signal-present images). One such radiograph is depicted in Figure 4.1 (a). The dimensions of the full-sized radiographs were 2048×2048 pixels and the cropped background images were of 420×420 pixels in size. A total of 234 background images were used in the experiments. Half of these were kept as the signal-absent images (normal clinical cases with no lung nodules).

In the next step, the simulated nodules (signals) were digitally superimposed on the centers of the remaining 117 images to create the set of signal-present images (abnormal clinical cases). The process of inserting the signals corresponds to that in [Samei et al., 1999]. The model of the signal used to simulate the lung nodules was developed by [Samei et al., 1997] and it mimics the radiographic characteristics of the tissue-equivalent lesions. The mathematical model of the contrast profile of the nodule was deduced from a database of real lung nodules [Samei et al., 1997]. It has been validated by means of a clinical study and is recognized by the medical imaging

community as realistic and acceptable [Samei et al., 2003]. Specifically, the contrast profile is defined as

$$c(|\mathbf{r}|) = C\left(\frac{4}{D^4}|\mathbf{r}|^4 + \frac{4.2}{D^2}|\mathbf{r}|^2 + 1\right),\ -0.6D \leq |\mathbf{r}| \leq 0.6D, \tag{4.1}$$

where $|\mathbf{r}|$ is the radial distance, C is the peak contrast value of the nodule, and D is the diameter of the nodule at the specified imaging plane. A typical value for the peak contrast-to-diameter ratio of a simulated nodule is 0.0098 mm^{-1}, the value found by [Samei et al., 1999] to correspond to spherical, uniform, muscle-equivalent lesions within the lungs. The profile of the signal is illustrated in Figure 4.2 (a). Conveniently, these nodules have a circular symmetry.

The specific values of the signal parameters in our study were guided by those from [Samei et al., 1999]. The diameter of the simulated nodule was $D = 30$ pixels. The nodule peak contrast C varied in a random fashion between the value 0.004948 and 0.05952, corresponding to the values between 1.2 and 3 JND units from the extension of the JNDmetrix model [Marchessoux et al., 2008b]. In common terms of contrast diameter product CD, the signals were in the range $CD = [0.026, 0.312]$. For illustration, the changes of pixel gray level values along the central line of an image are depicted in Figure 4.2 (b) and (c). The horizontal line along which we observe the pixel gray levels is chosen to avoid the surrounding variations caused by lung tissue or the ribs in the immediate neighborhood of the signal (see Figure 4.1 (c)).

As the final point in image data creation, the images were displayed on a medical three-mega-pixel grayscale display for radiography [Kimpe et al., 2007] and captured by a high resolution scientific camera at the resolution of 508×508 pixels. We used a ProMetric camera which allows the images to be taken in the XYZ color space (as defined by the Commission Internationale de l´éclairage, CIE). We used the Y component of the XYZ images which represents the luminance in cd/m^2.

Overall, the high complexity of the image preparation process implies a relatively small number of images in the study. For the purpose of MRMC studies, the camera captured luminance images of size 508×508 pixels were randomly split in two groups: one group of 72 image pairs for the trainer data set and another group of 45 image pairs for the tester data set (Figure 4.1). There was no overlap between the trainer and the tester data sets.

4.2.3.2 Observer performance experiments

The performance of the CHO is evaluated in MRMC studies, using the ssCHO from Section 3.3.1 as the observers (readers). The main principles of the model design and the usage of the trainer $\mathbf{g}^{\text{TR}}$ and tester images $\mathbf{g}$ are illustrated in Figure 4.3. First, in the training phase, we use the N_{tr} trainer image pairs (N_{tr} signal-absent and N_{tr} signal-present images) to train the model, *i.e.*, to estimate the CHO template $\mathbf{w}_{\text{CHO}}$. Next, in the testing phase, we apply that template on the N_{ts} tester image pairs (N_{ts} signal-absent and N_{ts} signal-present images) and compute the test statistic

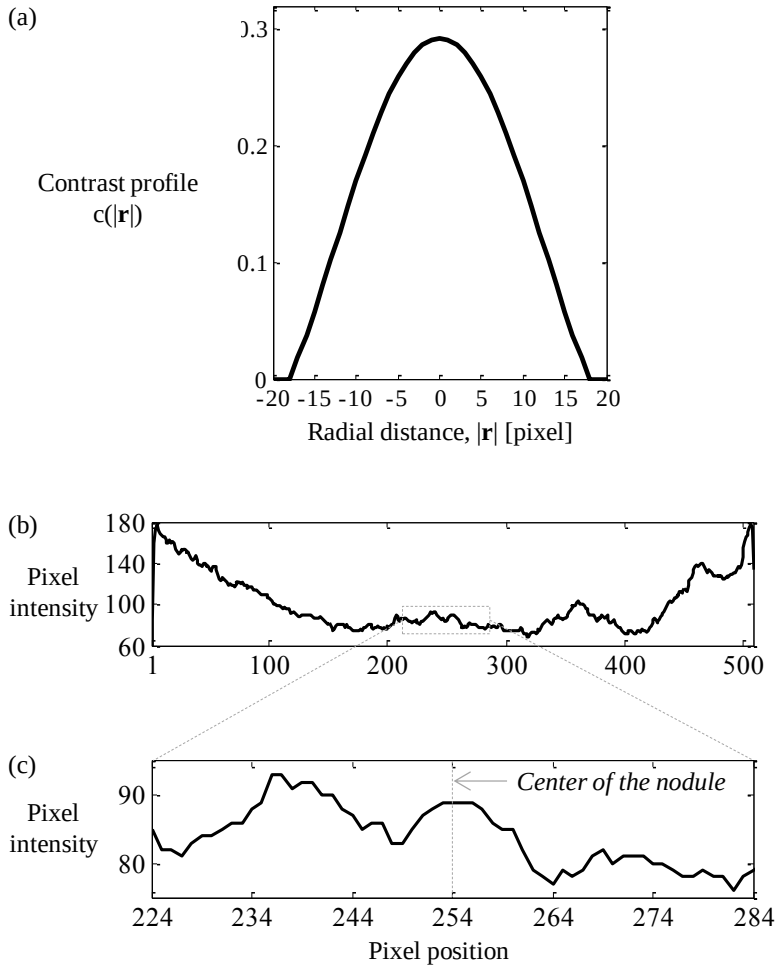


Figure 4.2: (a) Contrast profile of a simulated subtle lung nodule. The diameter of the nodule is $D = 30$ pixels and the peak contrast-to-diameter ratio of the nodule is 0.0098 mm^{-1}. (b) Pixel gray level values along the line indicated in Figure 4.1 (c). (c) Middle part of the graph (b) is enlarged to depict the pixel gray level values around the center of line where the nodule is located.

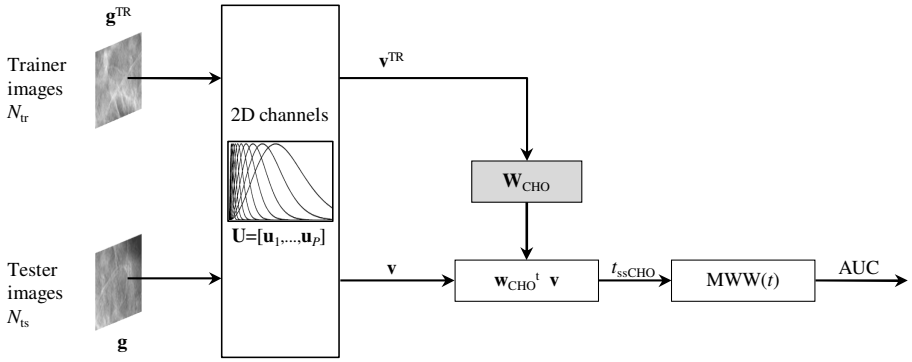


Figure 4.3: Flow chart for the single-slice channelized Hotelling observer model, ssCHO. The scheme is explained in the text. Further details can be found in Section 3.3.1.

("rating") for each tester image $\mathbf{g}$. Lastly, we calculate the Mann-Whitney-Wilcoxon (MWW) statistic to estimate the area under the ROC curve (AUC), the figure of merit for detection performance for the model.

As discussed in Section 4.1.1, an MRMC study configuration is determined by the readers and by the cases they read. Translated into the terms of a model observer study, the readers correspond to the estimated CHO templates (a different reader means a different (sub)set of the trainer images) and the cases are the tester images. In our study, every reader reads every case (fully-crossed MRMC study design), *i.e.*, all readers read exactly the same set of tester images.

The specific MRMC experiments are designed in light of the overall goal of the study to explore the effects of different parameters of a model observer study, related either to the model itself (in this case, the sscHO) or to the MRMC design. Specifically, we will investigate the impact of the following parameters: the ssCHO channel parameters, the number of trainer image pairs per reader, and the number of readers. Given the rather small number of available images in the experiment, we decide to give priority to the exploration of the effects associated with the number of trainer images, and keep the number of tester images fixed.

Taking into account the circular symmetry of the signal in our images, we choose the channels for the ssCHO to be the Laguerre-Gaussian (LG) polynomials. In line with the discussion from Section 3.5.2 in Chapter 3, we explore a range of different values for the number of LG channels, $P_{\mathrm{LG}} = \{5, 10, 15, 30\}$ (the first P_{LG} polynomials) as well as for the channel spread parameter $a_u = \{15, 25, 30, 45\}$.

The number of the trainer image pairs per reader is varied among $N_{\mathrm{tr}} = \{10, 25, 50\}$. Different readers are created using a different random selection of images from the trainer data set. Note here that because of the small number of the trainer images (a total of 72 image pairs), the readers are not always independent (*i.e.* some trainer

images may be used to train more than a single ssCHO reader). Consequently, the MRMC assumption of the reader independence in not valid in most of the experiments (test scenarios); this is discussed further in the results section.

The last explored parameter is the number of readers. It is varied among $N_{\mathrm{rd}} = \{5, 10, 25\}$. The other MRMC parameter, the number of tester image pairs, is kept fixed. The tester data set is the same, irrespective of other study parameters, and there is a total of $N_{\mathrm{ts}} = 45$ tester image pairs. With this, we are in fact assessing the effect of the number of readers on the variability of the AUC performance averaged across readers (*e.g.* by using the one-shot analysis [Gallas, 2006]).

The aforementioned parameters of interest (the four values of P_{LG}, the four values of a_u, the three values of N_{tr}, and the three values of N_{rd}) are varied in a factorial design. Thus, the total number of MRMC experiments in the study is $4 \times 4 \times 3 \times 3 = 144$.

As a standard methodology, the area under the ROC curve (AUC) is used as a figure of merit for the CHO performance [Barrett et al., 1998]. In addition, we compute the values of the detection signal-to-noise ratio (SNR) measure as defined by Eq. (3.16). Experimental AUC data is averaged across readers and the accuracy of the reader-averaged AUC is assessed using one-shot analysis [Gallas, 2006].

4.2.4 Results and discussion

The results of the 144 MRMC experiments are summarized in Table 4.2. The results in Table 4.2 are grouped by the LG channel parameter a_u, then by the number of LG channels P_{LG} and finally by the number of the trainer image pairs N_{tr}. The columns correspond to different numbers of MRMC readers N_{rd}. For each experiment, we show the one-shot estimates of the reader-averaged AUC value together with its corresponding standard deviation denoted by Std.[6] In each of the 144 experiments, all N_{rd} readers read exactly the same set of images comprised of the $N_{\mathrm{ts}} = 45$ tester image pairs. As indicated before, in order to not bias the ssCHO model predictions, the set of images used for the training of the model does not overlap with the set of tester images.

Because the total number of images in the study is rather small (due to the time consuming process of image preparation which is described in Section 4.2.3.1), so is the total number of the trainer image pairs $N_{\mathrm{tr}} = 72$. As a result, the reader independence assumption cannot hold in most of the test scenarios that we inspect, certainly not for the parameter configurations which assume $N_{\mathrm{rd}} = \{10, 25\}$. Therefore, error bars estimated by the one-shot analysis need to be interpreted with caution.

In addition, for the purpose of assessing the one-shot estimates of the AUC variance under the adequate condition of statistically independent readers, we perform an additional MRMC experiment consisting of 5 independent readers, each trained on

[6]The standard deviation of the AUC is the square root of the corresponding variance estimated by the one-shot method.

Table 4.2: One-shot analysis of the 144 model observer experiments

			$N_{\mathrm{rd}} = 5$		$N_{\mathrm{rd}} = 10$		$N_{\mathrm{rd}} = 25$	
a_u	P_{LG}	N_{tr}	AUC	Std	AUC	Std	AUC	Std
15	5	10	0.59	0.033	0.62	0.037	0.63	0.038
		25	0.62	0.054	0.64	0.052	0.63	0.052
		50	0.64	0.057	0.65	0.055	0.65	0.055
	10	10	0.65	0.031	0.66	0.041	0.63	0.034
		25	0.69	0.051	0.71	0.048	0.69	0.045
		50	0.73	0.049	0.74	0.049	0.74	0.049
	15	10	0.64	0.029	0.65	0.033	0.61	0.026
		25	0.67	0.045	0.70	0.041	0.70	0.040
		50	**0.77**	**0.047**	**0.78**	**0.047**	**0.77**	**0.047**
	30	10	0.54	0.044	0.52	0.048	0.52	0.045
		25	0.71	0.033	0.73	0.025	0.76	0.038
		50	*0.80*	*0.041*	*0.81*	*0.040*	*0.81*	*0.038*
25	5	10	0.72	0.045	0.72	0.042	0.72	0.039
		25	0.74	0.054	0.76	0.049	0.75	0.048
		50	**0.77**	**0.048**	**0.78**	**0.048**	**0.77**	**0.048**
	10	10	0.66	0.038	0.67	0.038	0.65	0.034
		25	0.70	0.051	0.71	0.048	0.69	0.047
		50	0.73	0.053	0.74	0.051	0.73	0.050
	15	10	0.65	0.039	0.60	0.036	0.60	0.025
		25	0.70	0.051	0.71	0.045	0.69	0.044
		50	0.75	0.047	0.76	0.047	0.75	0.048
	30	10	0.51	0.028	0.51	0.048	0.51	0.046
		25	0.64	0.047	0.63	0.035	0.62	0.035
		50	0.73	0.047	0.72	0.048	0.72	0.047
30	5	10	0.73	0.047	0.73	0.044	0.72	0.039
		25	0.74	0.053	0.75	0.048	0.74	0.047
		50	**0.77**	**0.049**	**0.77**	**0.048**	**0.77**	**0.048**
	10	10	0.66	0.042	0.67	0.038	0.65	0.030
		25	0.71	0.051	0.71	0.048	0.70	0.047
		50	0.75	0.050	0.75	0.050	0.74	0.050
	15	10	0.63	0.036	0.62	0.036	0.63	0.028
		25	0.69	0.047	0.70	0.045	0.68	0.045
		50	0.74	0.049	0.75	0.048	0.74	0.049
	30	10	0.51	0.028	0.51	0.048	0.51	0.046
		25	0.62	0.050	0.61	0.036	0.61	0.037
		50	0.71	0.047	0.72	0.047	0.72	0.048
45	5	10	0.74	0.042	0.74	0.043	0.73	0.039
		25	0.72	0.046	0.74	0.046	0.73	0.047
		50	0.76	0.049	0.76	0.048	0.76	0.049
	10	10	0.69	0.039	0.68	0.041	0.66	0.032
		25	0.70	0.053	0.70	0.049	0.69	0.047
		50	0.75	0.049	0.75	0.049	0.74	0.050
	15	10	0.66	0.052	0.60	0.042	0.58	0.024
		25	0.64	0.050	0.65	0.049	0.65	0.047
		50	0.72	0.053	0.71	0.052	0.70	0.053
	30	10	0.50	0.030	0.50	0.048	0.50	0.051
		25	0.63	0.042	0.63	0.033	0.62	0.036
		50	0.71	0.049	0.71	0.048	0.71	0.049

a separate non-overlapping set of 10 trainer image pairs; the LG channel parameters are $a_u = 25$ and $P_{\mathrm{LG}} = 5$. This experiment is repeated for 10 different selections of the trainer pairs (selection without replacement) and the obtained AUC variances are below 0.002 (Std< 0.045), which is in the range of the corresponding Std values in Table 4.2. This suggested that the one-shot estimates can be considered credible despite the readers are not completely independent.

4.2.4.1 Effects of the ssCHO training

First, we explore the trends in the AUC averaged across readers as the size of the trainer data set increases from $N_{\mathrm{tr}} = 10$ to $N_{\mathrm{tr}} = 50$ image pairs. Figure 4.4 compares the average AUC for $N_{\mathrm{tr}} = \{10, 25, 50\}$ in the case of $P_{\mathrm{LG}} = 5$. As expected, we observe that the range of the AUC values gradually increases with the increase of the number of the trainer images, from slightly below 0.75 for $N_{\mathrm{tr}} = 10$, to slightly over 0.75 for $N_{\mathrm{tr}} = 25$, to approximately 0.77 for $N_{\mathrm{tr}} = 50$. In line with that, as shown in Figure 4.5, the spread of the ROC points corresponding to the different readers is getting slightly narrower with increasing value of N_{tr} (this is a visual observation, not statistically confirmed). Further on, the uncertainty of the model observer performance due to the between-reader variability goes down as the number of the trainer images goes up. We remark here again that the size of data set available in this study is limited and in that sense it imposes certain limitations to the possible configurations of our MRMC studies.

Next, we refer to the values of the one-shot estimates of the standard deviation Std of the AUC. We observe the general trend of Std gradually increasing with the number of the trainer images. Moreover, Std gets less dependent on the number of readers as the size of the trainer data set grows. The latter is explained by the amount of information which is used to train the ssCHO, *i.e.*, the more trainer images per reader, the less the variability between readers. The exception to these conclusions are most of the study configurations with the number of LG channels set to 30. There, we observe larger variations in Std values for the different N_{tr} values. Especially, we notice that Std is highest in the MRMC configurations with the fewest trainer images, $N_{\mathrm{tr}} = 10$. Besides, the AUC values in these configurations are very low, close to 0.5. This is probably caused by the fact that 30 channels use more significant amount of image data to estimate the ssCHO template (train the ssCHO) than, for example, the 10 channels would use. However, the small size of the trainer image sample is not a reliable representative of its class which results in a low detection performance of the observer. Remember from Chapter 3 that the size of the covariance matrix which determines the ssCHO training is ${P_{\mathrm{LG}}}^2$. Thus, assuming $P_{\mathrm{LG}} = 30$, we would need at least 900 image samples to reliably estimate the ssCHO template.

Finally, for both the AUC measure and the one-shot Std, we notice that the differences between MRMC configurations of $N_{\mathrm{tr}} = 25$ and $N_{\mathrm{tr}} = 50$ are less significant than those observed for $N_{\mathrm{tr}} = 10$ and $N_{\mathrm{tr}} = 25$. This "saturation" in the performance level, already at such small size of the trainer image set, may be explained by the

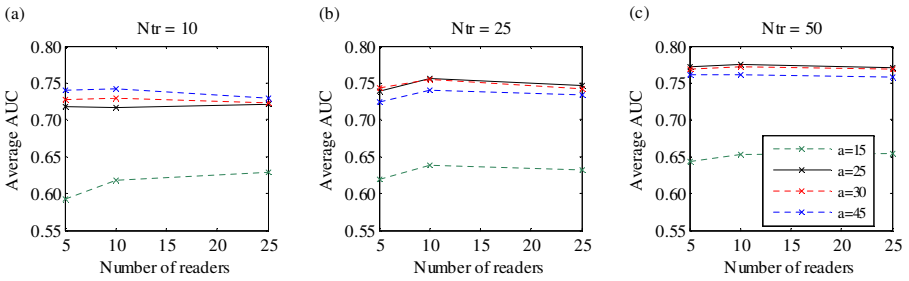


Figure 4.4: For $P_{\mathrm{LG}} = 5$, the AUC values averaged over $N_{\mathrm{rd}} = \{5, 10, 25\}$ readers. In each plot, different lines denote the value of the channel parameter $a_u = \{15, 25, 30, 45\}$. The three plots correspond to the three sizes of the trainer data set: (a) $N_{\mathrm{tr}} = 10$, (b) $N_{\mathrm{tr}} = 25$, (c) $N_{\mathrm{tr}} = 50$.

fact that the readers in these two test scenarios are significantly correlated due to the limited number of the available trainer images.

4.2.4.2 Exploring the LG channel parameters

Overall, the highest performance AUC ≈ 0.8 is observed in the experiment for which the parameters are set as follows: $a_u = 15$, $P_{\mathrm{LG}} = 5$, and $N_{\mathrm{tr}} = 50$ (see for the values marked in bold in Table 4.2). Nevertheless, as we already remarked, given the small number of the trainer images, these results should be treated with caution.

To select the best suited channel parameters for a given image data, we analyze the AUC values while observing the range of their estimated error bars ($\pm 2\,\mathrm{Std}$). Based on this criterion, we select two sets of ssCHO parameters. For the first ssCHO, the channel parameter is $a_u = 25$ and the number of LG channels is $P_{\mathrm{LG}} = 5$. For the second ssCHO, the corresponding values are $a_u = 30$ and $P_{\mathrm{LG}} = 5$. These two observers demonstrate quite similar performance of $\mathrm{AUC} = 0.77 \pm 2\,0.048$.

As a remark, we note that also another two configurations of the ssCHO parameters exhibit the same or even higher AUC performance, the ssCHO with $P_{\mathrm{LG}} = 30$ and $a_u = 15$, and the ssCHO with $P_{\mathrm{LG}} = 15$ and $a_u = 15$. However, due to the high variation in performance of the CHO with $P_{\mathrm{LG}} = 30$ which is discussed in the previous subsection, we keep this parameter out of the final selection. Similar applies for the CHO with $P_{\mathrm{LG}} = 15$ and $a_u = 15$.

To make the final decision about the LG channel parameters which are best suited for the given image data, we look in more detail at the performance of the two preselected channel configurations. In Figure 4.5 we visualize the ROC points for the corresponding MRMC experiments. Overall, the ssCHO with $a_u = 25$ seems to exhibit slightly more consistent performance (less variability) than the ssCHO with $a_u = 30$.

Based on the analysis, we chose the LG parameters of $P_{\mathrm{LG}} = 5$ and $a_u = 25$.

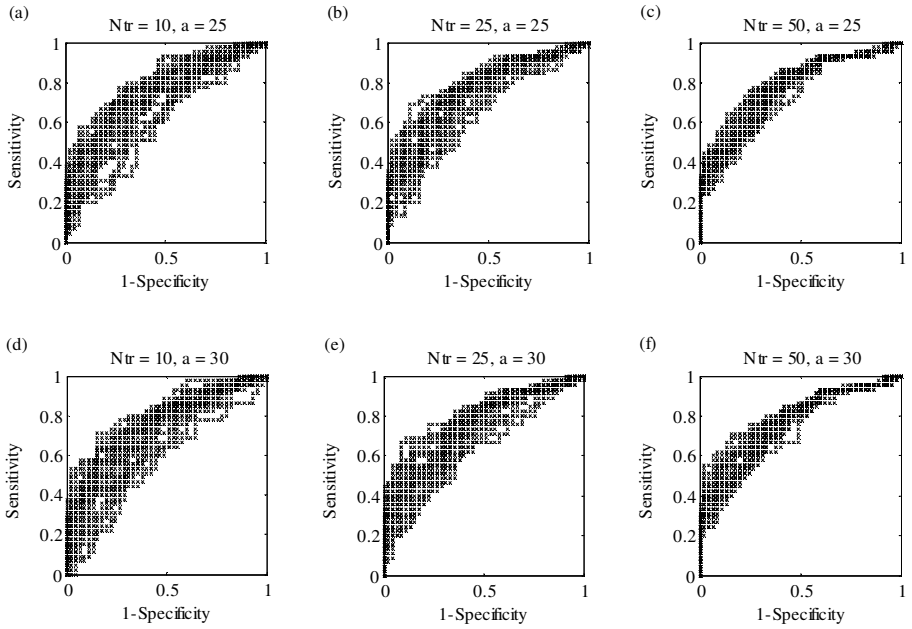


Figure 4.5: The spread of the ROC points for the two study experiments which demonstrate the highest performance in terms of AUC for $N_{\mathrm{tr}} = 50$. In both cases, the number of LG channels is $P_{\mathrm{LG}} = 5$ and the number of readers is $N_{\mathrm{rd}} = 25$. Top row: the value of LG channel parameter is $a_u = 25$ for different number of trainers: (a) $N_{\mathrm{tr}} = 10$, (b) $N_{\mathrm{tr}} = 25$, (c) $N_{\mathrm{tr}} = 50$. Bottom row: the value of channel parameter is $a_u = 30$ for different number of trainers: (d) $N_{\mathrm{tr}} = 10$, (e) $N_{\mathrm{tr}} = 25$, (f) $N_{\mathrm{tr}} = 50$.

This corresponds to the ssCHO performance of $\mathrm{AUC} = 0.77 \pm 2\,0.048$. Figure 4.6 illustrates the performance of the selected ssCHO observed for the different configurations defined by varying the number of readers $N_{\mathrm{rd}} = \{5, 10, 25\}$ and the number of trainers $N_{\mathrm{tr}} = \{10, 25, 50\}$.

4.2.4.3 Reflections on a related human observer study

Finally, in order to even roughly assess the potential of the ssCHO model for optimizing and evaluating medical displays for chest x-ray images, we refer to the human observer experiments in [Samei et al., 1999]. The authors use images similar to those in our study to evaluate the relative influence of quantum and anatomic noise on detectability of low-contrast subtle lung nodules in chest radiographs. Their results suggested a strong influence of anatomic noise (the anatomic structured pattern of the thorax) on the performance of humans while the effects of quantum noise were much less pronounced.

Specifically, when evaluating the effects of anatomic noise, [Samei et al., 1999]

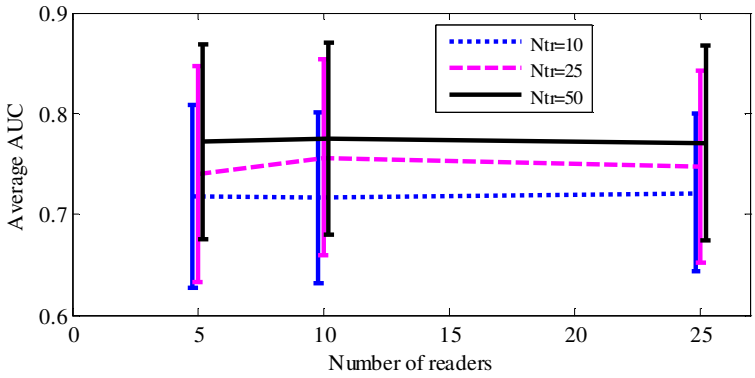


Figure 4.6: Average AUC of the ssCHO using $P_{\mathrm{LG}} = 5$ channels with the channel parameter $a_u = 25$ estimated for the nine study configurations defined by varying the number of readers $N_{\mathrm{tr}} = \{5, 10, 25\}$ and the number of trainer image pairs per reader $N_{\mathrm{ts}} = \{10, 25, 50\}$. Error bars are ± 2 standard deviations estimated by the one-shot method [Gallas, 2006].

consider clinical backgrounds (similar to those in our study) with three possible exactly known signals (SKE task). The signals are described by Eq. (4.1) where the product of the diameter and peak contrast of the signal take the values of $CD = \{0.14, 0.20, 0.28\}$. Depending on the signal parameters, the detection performance averaged over 5 human observers, was estimated at AUC ≈ 0.7 for $CD = 0.14$, AUC ≈ 0.8 for $CD = 0.20$, and AUC ≈ 0.9 for $CD = 0.28$. This AUC increase with CD was almost twice as slowly as in quantum noise (with statistical significance). That suggested that the dominant influence of anatomic noise on the detection of lung nodules.

In our ssCHO study, there was more variability in the signal parameters (the value of CD varied in a random fashion among the values of 0.026 and 0.312) which made the detection task more difficult. The estimated value of the AUC in our experiments was 0.77. Overall, considering the larger variability of the signal parameters in our study, the ssCHO performance seems to compare reasonably well with that of the human observers.

Certainly, a more detailed quantitative comparison between human and model observer performance would require the parameters of the image data to be further aligned between the two studies, both for the backgrounds (using images of exactly the same patients and acquired under exactly the same scanner parameters) and for the signals (using simulated signals of exactly the same CD values). Most importantly, careful attention should be taken to adequately incorporating the *effects of image display* in the simulation of images processed by the models.

Though not exactly for the case of the chest x-ray images, the model observer studies described further on in this chapter address exactly the issue of simulating the

effects of image display and demonstrate the importance of choosing or developing accurate models.

4.3 Reduced signal detectability due to the slow response time of a medical LCD

This is the first of three model observer studies which explore the impact of slow temporal response of a medical liquid crystal display (LCD) on signal detectability when the images are displayed/viewed in a browsing mode (moving through the sequence slice after slice). The other two studies are presented in Section 4.4 and Section 4.5.

4.3.1 Study rationale

In clinical practice today, planar imaging data sets are largely being replaced by volumetric ones. In parallel, cathode ray tube (CRT) devices are being replaced by the LCDs.[7] This trend is consistently observed in various anatomical as well as functional 3D imaging modalities including ultrasound, PET/SPECT, MRI and 3D breast imaging. Often, radiologists interpret these 3D images in a so-called *browsing* (or sequence-browsing, or stack-browsing) mode using a dedicated medical LCD where slices of a reconstructed volume are shown sequentially. However, despite the significantly improved quality of LCDs over the last few years, this technology still needs improvement in terms of the temporal response [Liang and Badano, 2007]. Here, *temporal response*, or *response time*, refers to the amount of time needed for the display pixel luminance to change from its current value (corresponding to the current slice of the displayed image) to the desired new value (corresponding to the subsequent slice of the displayed image). Due to the specifics of LCD technology, this transition from one luminance to another is not instantaneous. The details are discussed in Section 4.3.3.

Importantly, in a recent study of [Liang et al., 2008], the authors found that the slow response of LCD systems degrades the detection performance of model observers in the browsing mode of volumetric image reading by reducing the effective luminance contrast of the lesions. Specifically, they considered synthetic image sequences with a 2D signal superimposed on the central slice in the sequence. The display effects coming from the slow LCD response time were simulated using the temporal response model reported in [Liang and Badano, 2007].[8] In the model observer study of [Liang et al., 2008], two CHO designs were considered: the conventional 2D-CHO [Myers and Barrett, 1987], in this book referred to as the ssCHO, and that same CHO with

[7] In displaying static images, high performance LCD systems are considered to have comparable or better performance compared to the CRT devices. The only critical aspects in that case are noise [Badano et al., 2004] and viewing angle [Badano and Gallas, 2006].

[8] The images in our model observer study are exactly the same as in [Liang et al., 2008]. The details can be found in Section 4.3.4.2 and Section 4.3.4.1.

incorporated contrast sensitivity of the human visual system, a contrast-sensitive ss-CHO [Park et al., 2009a]. Each of the two ssCHO variants were used with only the central slice of the sequence (corresponding to the location of the signal, when it is present) rather than the whole sequence. As [Liang et al., 2008] pointed out, the limitation of such an approach was that the model observers which are designed for use in pure 2D detection tasks fail to incorporate 3D correlations in background and signals.

4.3.2 Experimental goal

Our objective is to evaluate the potential for using the multi-slice CHO (msCHO) designs proposed in Chapter 3 for the purpose of quantifying effects of the slow LCD response time when browsing volumetric images. The msCHO models are motivated by simplifying assumptions about humans browsing through a sequence of image slices, we refer to Section 3.3.3 for details. Simply given the fact that the msCHO gets more information about the image data then the ssCHO does, we expect the msCHO estimates of detectability to be more accurate (less pessimistic) than those of the ssCHO. Thus, first, we want to verify if the msCHO confirms what the ssCHO has suggested in [Liang et al., 2008]: the slow LCD response has a negative effect on signal detectability in browsing mode of 3D image viewing.[9] We assume here a pattern of moving from the first to the last slice in the sequence at a fixed speed. Next, we are interested in doing a comparative analysis of the ssCHO versus msCHO approach for both slow (30 fps) and fast (50 fps) browsing. Finally, the goal is to select the preferred msCHO design for the display investigations (out of the three types defined in Section 3.3.3: $\mathrm{msCHO_a}$, $\mathrm{msCHO_b}$, and $\mathrm{msCHO_c}$).

4.3.3 The basics of LCD temporal response simulations

Displays introduce certain "artifacts" to their input data. Therefore, in the context of image display-related investigations, it is important to distinguish between the image "before" (input to the display system) and "after" display (output of the display system). Hereafter, the image which is input to the display and used to drive the display is referred to as the *pre*-LCD image. When a pre-LCD image is shown on a display we refer to it as a *on*-LCD image. While in the previous Section 4.2 we used a high-end professional camera to capture the actual on-LCD images as they appear on the screen, in the studies presented here and in the following two sections (Section 4.4 and Section 4.5) we will use models to simulate LCD effects. In particular, we are only interested in the effects of the slow LCD response time (coupled with the display calibration); other LCD effects are not considered.[10]

[9] At the time of our study, there were still no literature reports of a related human observer study to confirm or reject the assumption that the slow LCD temporal response has a negative effect on signal detectability in the browsing mode of image reading. To our knowledge, the first such study to appear was the work of [Badano, 2009].

[10] Next to the response time and calibration of the display devices, a range of other parameters may affect the quality of displayed images, including but not limited to: spatial and temporal dithering (implemented

Before we move on to the basics of the temporal response model of an LCD display, let us first introduce the terms and notations used in this thesis to refer to on-LCD images. As mentioned earlier, we assume that a pre-LCD image sequence is comprised of N image slices, or *frames*. The pre-LCD images are viewed on a display device characterized by its *refresh rate* which determines the number of times that the screen is redrawn during 1 sec. As is common, the *frame rate* or *browsing rate*, f_{frame}, determines the number of frames that is displayed during 1 second. Hence, each frame is displayed during the time interval of $1/f_{\text{frame}}$ referred to as the *frame duration*, T_{frame}. Depending on the refresh rate of the display f_{refresh}, each frame from the image sequence is drawn on the screen one or more times within a frame duration. That is to say, the maximum frame rate is determined by the display technology, $f_{\text{frame}} \leq f_{\text{refresh}}$. Using $T_{\text{refresh}} = 1/f_{\text{refresh}}$ to denote the *refresh time* of the display, *i.e.*, the time between the two consecutive display refresh cycles, we can write the following

$$T_{\text{frame}} = \frac{f_{\text{refresh}}}{f_{\text{frame}}} T_{\text{refresh}}. \tag{4.2}$$

Here, we assume that the frame duration is constant over all frames in the image and thus each frame is displayed exactly $(T_{\text{frame}}/T_{\text{refresh}})$ times during the frame duration. Further on, we will use the term *frame repeat*, denoted by FR, to refer to the number of consecutive repetitions of a given frame in a on-LCD image, *i.e.*, the number of frames displayed per frame duration,

$$\text{FR} = \frac{T_{\text{frame}}}{T_{\text{refresh}}} = \frac{f_{\text{refresh}}}{f_{\text{frame}}}. \tag{4.3}$$

For example, let us consider a display with $f_{\text{refresh}} = 50$ Hz ($T_{\text{refresh}} = 20$ ms) when the frame rate is $f_{\text{frame}} = 25$ frames per second (fps). Given Eq.(4.2), the frame duration is $T_{\text{frame}} = (50/25)20 = 40$ ms. And, in line with Eq.(4.3), each slice is displayed exactly $\text{FR} = 40/20 = 2$ times during T_{frame}. This considered, a straight forward simulation of the display (still not including any effects of the slow LCD response time) would result in a on-LCD image sequence that consists of a total of $2N$ frames: $\text{FR} = 2$ times as many frames as there are in the corresponding pre-LCD image. If we denote the n-th frame of the pre-LCD image sequence by $\mathbf{g}^0_{(n)}$, $n = 1, \ldots, N$, then the corresponding on-LCD sequence at $\text{FR} = 2$ can be described as $\mathbf{g} = \left[\mathbf{g}^0_{(1)}, \mathbf{g}^0_{(1)}, \mathbf{g}^0_{(2)}, \mathbf{g}^0_{(2)}, \ldots, \mathbf{g}^0_{(N)}, \mathbf{g}^0_{(N)}\right]$. In words, the first FR frames of the on-LCD sequence equal the first frame of the pre-LCD frame, the next FR on-LCD frames equal the second pre-LCD frame, and so on until the last FR on-LCD frames which equal the last pre-LCD frame. Likewise, a on-LCD image in the case of $\text{FR} = 3$

by manufacturers to achieve a more precise calibration), spatial noise (stationary differences in the behavior of individual pixels), and viewing angle dependency (the luminance and contrast properties of LCDs depend on the angle from which the display face is observed). Details of the complete display modelling within the MEVIC simulation platform can be found in [Marchessoux et al., 2008a] (for LCD) and in [Marchessoux and Jung, 2006] (for CRT devices). Other relevant references from the literature include [Badano et al., 2003, Kimpe et al., 2005, Samei et al., 2005, Fetterly et al., 2008].

($f_{\text{frame}} = (50/3) \approx 17$ fps) would comprise a total of $3N$ frames (see Figure 4.7), and so forth.

Now, we discuss the issue of slow LCD temporal response and explain the difference between the pixel intensity values in pre- and on-LCD images. Let us denote $g(x, y, n)$ the gray level of the image pixel at position (x, y) within frame n, where $x = 1, ..., X$, $y = 1, ..., Y$ and $n = 1, ..., N$. The gray levels vary in the range from 0 to $g_{\max}$ (in this study, the images are 8-bit and so $g_{\max} = 256$). Similarly, we use $l(x, y, n)$ to denote the level of luminance corresponding to $g(x, y, n)$. Here, the grayscale pixel intensity is considered an input to the display (the digital drive level, ddl) and the output is quantified as the luminance of the corresponding display pixel(s). The mapping from gray level to luminance and vice versa is implemented using the measured *luminance curve* of the display, $c(g)$ (see for example Figure 4.9 (a)). For simplicity, we assume the images are displayed in native resolution, *i.e.*, each pixel in the input image slice matches exactly one pixel on the display. Thus, further on we consider a single pixel at position (x, y) within a given frame and drop the corresponding position indices, *i.e.*, we use $g(n)$ to mean $g(x, y, n)$ and $l(n)$ to mean $l(x, y, n)$.

In general, we assume that pixel intensity values in the pre-LCD image sequence change from frame to frame, from $g^0(n)$ in the current frame to $g^0(n + 1)$ in the subsequent frame. Ideally, when the response time of the display would be zero (infinitely fast LCD), the corresponding luminance of the display pixel would exactly match the input grayscale values and it would be approximately constant throughout the frame duration. Accordingly, transitioning from frame n to the frame $(n + 1)$, the luminance of the display pixel would instantaneously change from $l^0(n) = c\left(g^0(n)\right)$ into $l^0(n + 1) = c\left(g^0(n + 1)\right)$. Indeed, in the case of static images or very low frame rates (pausing after each frame), we can assume that T_{frame} is long enough that the aforementioned ideal-case assumptions may hold true. However, due to non-ideal temporal response of the LCD (large reorientation times of the liquid crystal cells), the transition of luminance from one frame to another in not instantaneous. This is exactly the effect of the LCD displays which we are interested in modelling.[11]

The following explanation is supported by Figure 4.7. Let $l(n, T_{\text{frame}})$ denote the luminance level of a given pixel achieved at the end of the reference frame n. In the new frame $(n+1)$, the target luminance level of the exact same pixel on the display is $l^0(n+1) = c(g^0(n+1))$ where $l^0(n+1) \neq l(n, T_{\text{frame}})$. Ideally, we would have that $l(n + 1, t) = l^0(n + 1)$, $t \in (0, T_{\text{frame}}]$. Nonetheless, depending on the magnitude of the target luminance transition $\triangle l(n+1, n) = l^0(n+1) - l(n, T_{\text{frame}})$, and depending on the LCD response time for that transition, it may take multiple display refresh intervals T_{refresh} for a given liquid crystal element to achieve the target luminance level $l^0(n+1)$ and thus complete the target transition $\triangle l(n+1, n)$; this is also referred to as the "trailing effect". Moreover, when the frame duration T_{frame} is smaller than

[11] Note that the LCD models investigated in this chapter address solely the effect of the slow LCD response time (coupled with the display calibration characterized by the luminance response curve $c(g)$) and no other effect of LCD image display.

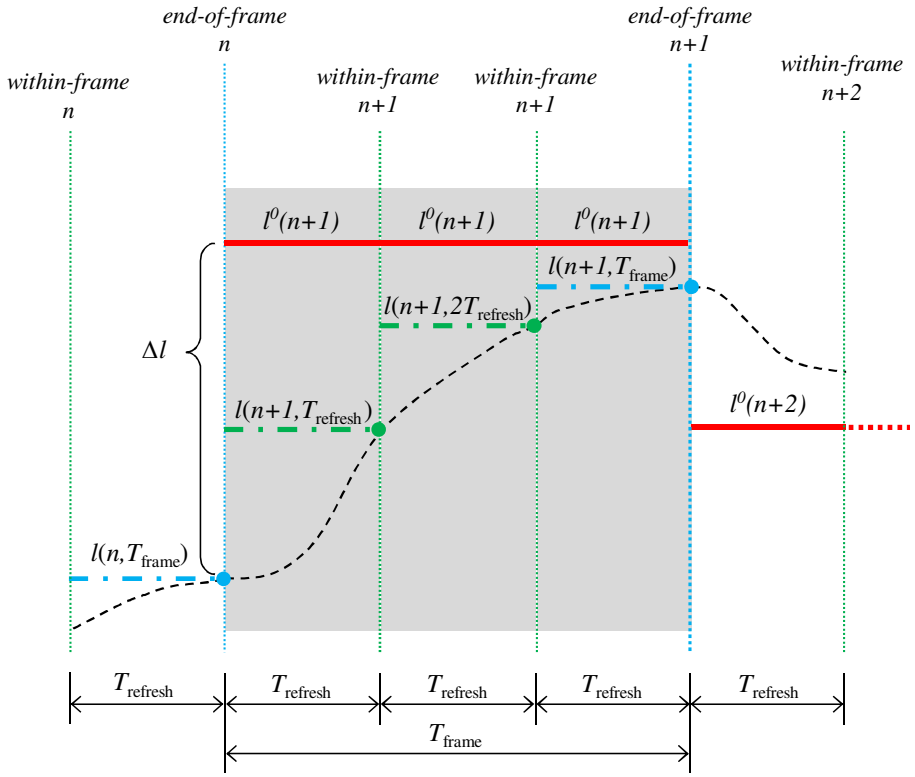


Figure 4.7: Pixel luminance changes at subsequent on-LCD frames. $l(n, T_{\text{frame}})$ is the achieved luminance of a given pixel at the end of frame n. For the following frame $(n+1)$, the target luminance level is $l^0(n+1) = c\left(g^0(n+1)\right)$, where $c(g)$ is the the luminance curve of the given display (see Figure 4.9 (a) for an example). Because of the slow LCD response time, the target luminance transition $\triangle l(n+1, n) = l^0(n+1) - l(n, T_{\text{frame}})$ is not achieved instantaneously (the trailing effect, see text). In fact, the luminance changes gradually over the frame duration T_{frame}, from $l(n+1, T_{\text{refresh}})$, over $l(n+1, 2T_{\text{refresh}})$, up to the level of $l(n+1, 3T_{\text{refresh}}) = l(n+1, T_{\text{frame}})$. Note that in this example even at the end of the frame $(n+1)$ the target luminance has not been achieved, *i.e.*, $l(n+1, T_{\text{frame}}) < l^0(n+1)$. Further on, during the time of the subsequent frame $(n+2)$, the target luminance transition is $\triangle l(n+2, n+1) = l^0(n+2) - l(n+1, T_{\text{frame}})$ where $l^0(n+2) = c\left(g^0(n+2)\right)$.

the corresponding response time of the LCD cell, the target luminance level $\triangle l(n+1, n)$ will even not be achieved at the point of moving to the subsequent frame $(n+2)$. This is the case in the example from Figure 4.7 where at the time instance $t = T_{\text{refresh}}$ only less than half of the transition is completed, then at $t = 2T_{\text{refresh}}$ and $t = 3T_{\text{refresh}}$ the luminance gets closer to the target, but it does not get completed even by the end of the frame duration $T_{\text{frame}} = 3T_{\text{refresh}}$, *i.e.*, $l(n+1, T_{\text{frame}}) < l^0(n+1)$. Importantly, these "delayed" or "incomplete" luminance transitions result in the reduced effective luminance contrast of details in medical images (as will be seen in the experiments later on).

For the purpose of further analysis, we introduce also the terms of *within-frame* and *end-of-frame* luminance values denoted by l_{in} and l_{end}, respectively. Here, l_{end} is the luminance achieved at the end of each frame duration, $t = T_{\text{frame}}$ and l_{in} refers to the luminance values achieved at the end of the display refresh cycles within a given frame duration, *i.e.*, at time instances for which $t = kT_{\text{refresh}} < T_{\text{frame}}$, $k \in \mathcal{N}$. In the example from Figure 4.7, the end-of-frame luminance values occur at the time instances corresponding to each $T_{\text{frame}} = 3T_{\text{refresh}}$ while the within-frame luminance values can be measured at the time instances corresponding to T_{refresh} and $2T_{\text{refresh}}$.

Lastly, we describe the steps of the LCD temporal response simulations. To start with, the model parameters are the *luminance response curve* $c(g)$, which describes the mapping between luminance and grayscale values of the pixels (see Figure 4.9 (a) and Figure 4.14 (a)), and the matrix denoted by $\mathbf{Q}$ which characterizes *response times* for all possible grayscale (luminance) transitions for a given display device. Usually, the values of the matrix $\mathbf{Q}$ are collected from empirical measurements. For example, [Liang and Badano, 2007] measured the percentage of the target luminance transition reached after one frame at a given frame rate (see Figure 4.9 (b)). Alternatively, as proposed by [Marchessoux et al., 2008a], we could measure the time to complete the target transition (see Figure 4.14 (b)). Next, as input and output for the model, respectively, we define the pre-LCD grayscale image sequence $\mathbf{g}^0$ comprised of N slices (frames) each with $X \times Y$ pixels and the on-LCD luminance image $\mathbf{l}$ (and/or the grayscale image $\mathbf{g} = \text{nint}\, c^{-1}(\mathbf{l})$), where $\text{nint}(\cdot)$ denotes the closest integer to x.[12] Remember from the earlier descriptions in this section that the number of slices in the on-LCD image sequence can differ from the number of slices in the pre-LCD image, depending on the specifics of the display simulation.

For simplicity, we assume no display effects for the first slice in the sequence, or $l(1,t) = l^0(1) = c(g^0(1))$ where $t \in (0, T_{\text{frame}}]$. Further on, for slices at position $n = 2, \ldots, N$, the following computations are performed:

$$l^0(n) = c\left(g^0(n)\right), \tag{4.4}$$

$$\triangle l(n+1, n) = l^0(n+1) - l(n, T_{\text{frame}}), \tag{4.5}$$

$$l(n+1, t) = l(n, T_{\text{frame}}) + \rho\left(t, q(n+1, n)\right) \triangle l(n+1, n), \tag{4.6}$$

[12] Commonly, the luminance curve $l = c(g)$ maps luminance values to a floating-point representation of the grayscale values, while they are in fact the integer values in the range from 0 to $g_{\max}$.

where $t \in (0, T_{\text{frame}}]$. Here, $l(n, T_{\text{frame}})$ is the value of luminance achieved at the end of frame n, $l^0(n)$ is the target luminance value for the subsequent frame $(n+1)$, and $g^0(n)$ is its corresponding input grayscale value. The value denoted $\rho\,(t, q(n+1, n))$ specifies the completeness of the target luminance transition $\triangle l(n+1, n)$ at time instance t during the $(n+1)$ frame duration. The value of $q(n+1, n)$ is determined from the matrix $\mathbf{Q}$ as the value of the matrix element in column $g(n, T_{\text{frame}})$ ("From") and row $g^0(n)$ ("To"). As mentioned previously, the matrix $\mathbf{Q}$ can characterize different temporal attributes of the display (*e.g.* the percentage of $\triangle l$ reached after one frame, or the time to complete the target transition). In line with this, also the definition of ρ may vary. We will describe the details of the two different models used in our experiments later, in Section 4.3.4.2 and Section 4.4.3.2. As remarked earlier, the model is the same irrespective of the pixel position within the slice.

4.3.4 Study design and methodology

Our study design consists of (1) the generation of pre-LCD image data, (2) the simulations of the LCD temporal response effects, and (3) model observer experiments. In the following subsections we discuss each in more detail.

4.3.4.1 pre-LCD image data

We simulated a total of 2200 volumetric images of $256 \times 256 \times 64$ voxels in size, where $N = 64$ is the number of slices in the image sequence, each slice of the width $X = 256$ and height $Y = 256$. Thus, the total number of voxels in the image was $M = XYN = 4194304$.

The images were the same as in the previously mentioned ssCHO study of [Liang et al., 2008]. The background data were synthesized as 3D clustered-lumpy backgrounds (CLB) with mean number of clusters $K = 160$, mean number of blobs per cluster $N_k = 20$, $k = 1, \ldots, K$, and characteristic lengths $L_x = 3$, $L_y = 2$, $L_z = 3$. See Section 3.2.1 for the details about the CLB model. In their 2D version, CLBs have been shown to well mimic mammographic anatomical structure in their appearance [Bochud et al., 1999]. An example background slice is depicted in Figure 4.8 (b). The image data were grayscale in 8-bit integer precision where the average gray level of each image was 32 and the maximum gray level was 64. These pixel intensity parameters were chosen to correspond to the fastest LCD luminance transitions (according to the display measurements) and will be discussed further shortly, in Section 4.3.4.2. The details can be found in [Liang et al., 2008].

Half of the backgrounds were used as signal-absent images. To generate signal-present images, we added a 2D designer nodule signal [Burgess et al., 2001] in the central slice of the remaining 1100 backgrounds. This signal model was aimed to represent a simplified one-slice thick breast mass lesion. As illustrated in Figure 4.8 (a), three different peak intensities of the signal were considered, $a_s = \{4, 8, 16\}$. These a_s values were chosen to correspond to the ssCHO detectability range of 0.75 to 1

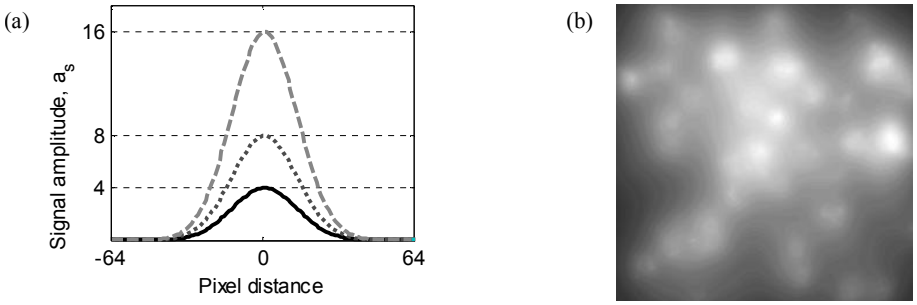


Figure 4.8: (a) Contrast profiles of the simulated signal designer nodules for three different amplitude values, $a_s = \{4, 8, 16\}$. (b) Central slice from an example signal-absent image volume of $256 \times 256 \times 64$ voxels. The image is a 3D CLB background object from Section 3.2.1 with the following parameters: mean number of clusters $K = 160$, mean number of blobs per cluster $N_k = 20$, $L_x = 3$, $L_y = 2$, $L_z = 3$.

in approximately equal steps (note that still no display effects were considered at this point). Each of the three signal amplitudes characterizes one contrast category with a total of 1100 pairs of signal-absent and signal-present images. They were used in the model observer experiments as representatives of the pre-LCD image sequences which do not take into account the temporal response of the display.

4.3.4.2 LCD temporal response simulations

The effects of LCD temporal response are simulated following the steps described in Section 4.3.3 while using the model parameters from [Liang and Badano, 2007]. In particular, we consider the display measurements associated with a five-mega-pixel medical color LCD depicted in Figure 4.9.

First, we use the luminance response curve from Figure 4.9 (a) to convert the pre-LCD images in grayscale to those in luminance space, $l = c(g)$. Then, in order to estimate the luminance of the pixel as seen on the screen – the on-LCD pixel luminance, we use the matrix $\mathbf{Q}$ from Figure 4.9 (b). In the measurements of [Liang and Badano, 2007], the frame rate is fixed at 30 fps ($T_{\text{frame}} = 1/30 \approx 33$ ms) and the elements $q(n+1, n) \in \mathbf{Q}$ are the percentage of the corresponding target luminance transition $\triangle l(n+1, n) = l^0(n+1) - l(n, T_{\text{frame}})$ reached after one frame duration. That is to say, the elements of $\mathbf{Q}$ correspond to the variable ρ from Eq. (4.6) at the time instance $T^* = 1/30 \approx 33$ ms. It is of interest to note from Figure 4.9 (b) that only a very minor portion of all possible luminance transitions actually get completed during T^* (look for the red color in the image of matrix $\mathbf{Q}$). Importantly for medical imaging, the fastest transitions seem to be those "to" either very low or very high luminance values (*i.e.* to very dark or nearly white grayscale values), almost independent of the "from" pixel intensity value. However, such transitions are not typical of medical image data. Conversely, the transitions in medical images are often in the

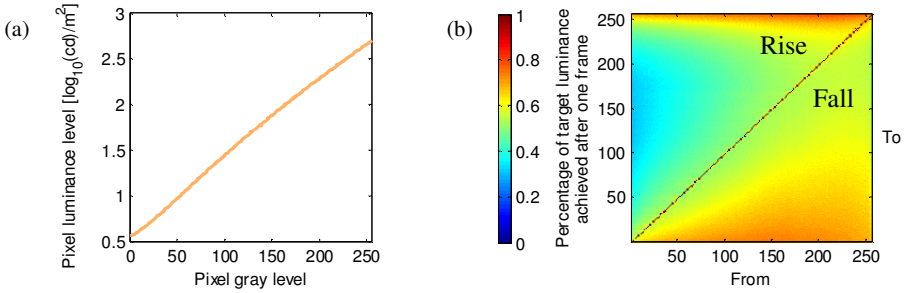


Figure 4.9: Parameters of the LCD temporal response model for the five-mega-pixel medical LCD from the study of [Liang and Badano, 2007]. (a) Luminance response curve of the display, $c(g)$. (b) Matrix $\mathbf{Q}$ of the percentage of the target luminance transition reached after one frame, measured at the frame rate of 30 fps. Here, the target luminance transition starts at the luminance determined by the x-coordinate of the matrix $l_{\text{from}} = c(g_x)$ and it ends at the luminance level determined by the y-coordinate of the matrix $l_{\text{to}} = c(g_y)$.

range of medium intensities which seem to remain incomplete after the frame duration. Even more worrying, the report of [Liang and Badano, 2007] suggests that a great number of the transitions would reach less than 80% even after four frame durations ($t = 4\,T^*$). The details can be found in [Liang and Badano, 2007] and [Liang et al., 2008].

For estimating the values of on-LCD luminance at the time instances other than $t = T^*$, [Liang and Badano, 2007] assume that the change of luminance is linear over time (for simplicity) and define the index of completeness of the target luminance transition as

$$\rho\left(t, q\left(n+1, n\right)\right) = \frac{1}{T^*}\,\rho\left(T^*, q\left(n+1, n\right)\right)\,t. \tag{4.7}$$

Irrespective of the frame rate, the model of [Liang and Badano, 2007] only computes the end-of-frame luminance values; the within-frame values are not computed (refer to Figure 4.7 for an illustration). Hence, the number of slices in the on-LCD image is exactly the same as that of the pre-LCD sequence. We discuss the consequences of this simplification later in Section 4.5.

For our experiments, two different frame rates are considered: a low frame rate of 30 fps and a high rate of 50 fps. The corresponding on-LCD data sets are named on-LCD-30 and on-LCD-50, respectively. Given the three pre-LCD image categories of the three different signal amplitudes, we end up with a total of $3 \times 3 = 9$ categories of simulated images.

4.3.4.3 Observer performance experiments

One of our goals mentioned in Section 4.3.2 is to compare the signal detection performance of the ssCHO versus msCHO at different frame rates. Therefore, the first

model observer in our experiments is the ssCHO described in Section 3.3.1. Since the signal in our study is restricted to one-slice only (slice $n = 32$, see Figure 4.11), the ssCHO is calculated for exactly that slice position in the sequence.

Another goal is to select the preferred msCHO design for the display-related model observer studies. From our experimental results in Chapter 3, the model $\mathrm{msCHO_b}$ seems least affected by the number of training samples. Depending on the contrast of the signal, the $\mathrm{msCHO_a}$ model may require some more training data to achieve the same performance as $\mathrm{msCHO_b}$, while the model $\mathrm{msCHO_c}$ requires notably more training data.[13] Given the timely display simulations, we exclude model $\mathrm{msCHO_c}$ from our analysis and we continue with two msCHO designs, $\mathrm{msCHO_a}$ and $\mathrm{msCHO_b}$.

We refer to Figure 4.10 for a brief overview of the model observer process; the details can be found in Chapter 3. Since a model observer is essentially a classifier, it has to be trained on the basis of training data for which we know the true class membership (signal-absent or signal-present). We use the superscript TR to denote the data associated exclusively with the training process. In terms of the considered model observers, training means estimating the CHO template of the ssCHO model, or the CHO and the HO templates of the msCHO models. Once the observer has been trained, it is applied on another set of images (again with a known class membership) and the results are used to assess the performance of the model. The specifics are reviewed next.

In the first stage, the observer processes the image sequence in planar view (xy-plane), slice after slice. A filter bank of 2D channels, $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_P]$ is applied on the image pixel values of each slice, $\mathbf{g}_{(r)}$, $r = 1, ..., R$. Here, P denotes the number of channels and R denotes the number of adjacent slices in the region of interest (ROI) in which the signal is located. As a result, we get the channelized slice data $\mathbf{v}_{(r)} = \mathbf{U}^t\mathbf{g}_{(r)}$. The channels in this study are the first ten dense difference-of-Gaussian (DDOG) channels shown in [Abbey and Barrett, 2001] to closely track human observer performance. Next, we use the channelized data to estimate the 2D-CHO template for each slice position, $\mathbf{w}_{\mathrm{CHO}}(r)$ and subsequently build the slice test statistics, $t_{(r)}$. It is at this step that the two considered msCHO variants differ. The $\mathrm{msCHO_a}$ design actually estimates a separate 2D-CHO template for each slice position r. By contrast, the $\mathrm{msCHO_b}$ design estimates only a single 2D-CHO template for the slice position in which the signal (dominantly) resides and uses that same template for all slice positions in the ROI. In the case of our images, this template is $\mathbf{w}_{\mathrm{CHO}}(R/2+1)$. Because the signal in our study is present in one slice only, we select

[13] Remember from Section 3.3.3.4 that the size of the data covariance matrix of $\mathrm{msCHO_c}$ model can be a critical factor in the cases where the available trainer data set is limited in size. Even if the number of image slices of interest is small, *e.g.*, $N = 5$ slices, and the number of channels in the model takes a typical value of $P = 10$, the number of elements in the covariance matrix of $\mathrm{msCHO_c}$ is quite large, $(N \times P)^2 = 2500$. This means that, in order to reliably train some $N_{\mathrm{rd}} = 5$ $\mathrm{msCHO_c}$ readers, we would need at least $2500 \times 5 = 12500$ trainer images. This is a pretty heavy requirement for the studies which involve the timely display simulations. Instead, it is of interest to consider alternative approaches to inversion of the estimated covariance matrix discussed in Section 3.3.3.4, *e.g.*, using regularized inversion as in [Zhang et al., 2013].

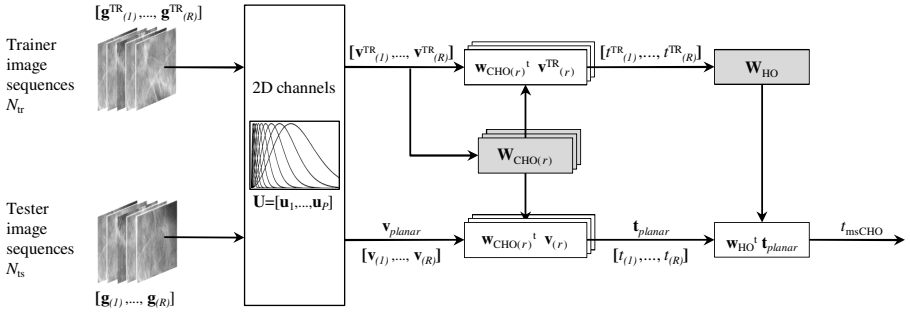


Figure 4.10: Overview of the model observer process. Assuming $r > 1$, the diagram correspond to the multi-slice models $\mathrm{msCHO_a}$ and $\mathrm{msCHO_b}$. The two models differ in how they estimate the templates $\mathbf{w}_{\mathrm{CHO}(r)}$, where $r = 1, ..., R$ identifies the position of the slice within the ROI. The model $\mathrm{msCHO_a}$ assigns a separate template to each slice position, $\mathbf{w}_{\mathrm{CHO}(r)}$. On the other hand, the $\mathrm{msCHO_b}$ works with a single template, *i.e.*, $\mathbf{w}_{\mathrm{CHO}(r)} = \mathbf{w}_{\mathrm{CHO}(R/2+1)}$ where $r = (R/2 + 1)$ denotes the position of the slice in which the signal is located. The two models are defined in Section 3.3.3.2 and Section 3.3.3.3, respectively. In the special case where $r = 1$, the diagram corresponds to the ssCHO model (see also Figure 4.3).

fewer slices for the ROI. Especially, we explore the ROIs of 3, 5, or 7 adjacent slices centered around the slice $n = 32$ in which the signal is located. The important aspect concerning the preferred number of slices in the ROI are discussed in Section 3.3.3.5. Note that in the special case of $r = 1$, the described process corresponds to the ssCHO model (see also Figure 4.3).

In the second stage, assuming $r > 1$, the msCHO integrates the information in the z-direction to result in the final test statistic ("rating") for the whole image sequence, t_{msCHO}. To do this, we first estimate the template $\mathbf{w}_{\mathrm{HO}}$ of the Hotelling observer (HO) [Barrett and Myers, 2004] using the training data $[{t_{(1)}}^{\mathrm{TR}}, \ldots, {t_{(R)}}^{\mathrm{TR}}]$. Last, we apply the $\mathbf{w}_{\mathrm{HO}}$ on the corresponding test data $[t_{(1)}, \ldots, t_{(R)}]$ and as a result we get the test statistic for the sequence t_{msCHO}. Eventually, the t_{msCHO} ratings of all test images are used to compute the MWW statistic and estimate the AUC figure-of-merit for the model.

Our model observer experiments comply with the paradigm of a fully-crossed MRMC design. As described in Section 4.3.4.1, a total of 9 image categories, each of 1100 pairs of one signal-present and another signal-absent image were generated. In the MRMC experiments, each set of 1100 image pairs is split in the following way: 1000 pairs are used as trainer data and 100 pairs are used as tester data. There, each trainer data set of 1100 image pairs is divided in 5 independent subsets of $N_{\mathrm{tr}} = 200$ pairs and used to train $N_{\mathrm{rd}} = 5$ readers. All readers read the same set of $N_{\mathrm{ts}} = 100$ tester image pairs. The models are compared in terms of their AUC values. For variance analysis, we use the one-shot method [Gallas, 2006].

4.3.5 Results and discussion

We analyze and discuss the following two effects of the LCD temporal response: (1) the effect on image data (how the on-LCD data differs from its pre-LCD input depending on the details of the LCD temporal response) and (2) the effect on model observer performance (how the signal detectability changes with increasing speed of sequence-browsing). Moreover, we look at (3) how the size of the ROI (number of successive slices) used in msCHO experiments affects the observer performance as the browsing speed increases. Finally, by comparing $\mathrm{msCHO_a}$ and $\mathrm{msCHO_b}$ performance to the human performance measured in a related human observer study from the literature [Badano, 2009], we give some (4) considerations on the preferred msCHO design for a given application.

4.3.5.1 Simulated on-LCD image data

The plots in Figure 4.11 depict the intensity profile of the central pixel in the slice (xy-plane) as we browse through the sequence. The changes in intensity are shown for an example pre-LCD image sequence and for its corresponding on-LCD images at the frame rates of 30 fps (on-LCD-30) and 50 fps (on-LCD-50). The three plots correspond to the three different contrasts of the signal: (a) $a_s = 4$, (b) $a_s = 8$, and (c) $a_s = 16$. Remember that the signal resides in slice $n = 32$ of the pre-LCD images. Looking at the intensity profiles of the on-LCD images, we see that at slice $n = 32$ the signal achieves only a fraction of its pre-LCD magnitude. In the case of $a_s = 8$ and $a_s = 16$, the intensity continues to increase also in the following slice $n = 33$ but it still fails to reach the peak value of the pre-LCD signal. On the other hand, once the signal stops increasing, it does not immediately disappear in the subsequent slice (as it is the case in pre-LCD image). Rather, the signal remains present over a few upcoming slices (slice $n = 34$ or even further) while gradually decreasing in intensity – the "trailing effect" caused by slow temporal response of the display. The spread of the trailing depends on both the signal intensity and the browsing speed: the higher the peak of the pre-LCD signal and the higher the frame rate, the thicker and longer the tails. In the example from Figure 4.11, we see the most pronounced trailing effect in the case of signal contrast $a_s = 16$ at the frame rate of 50 fps in Figure 4.11 (c). The signal expands from slice $n = 32$ all the way to slice $n = 36$ while also the peak of the signal moves from slice $n = 32$ to slice $n = 33$. Clearly, the on-LCD signal is largely different from the true pre-LCD signal. Lastly, note also the difference in pre-LCD versus on-LCD luminance in the signal-free image regions. While they may appear less severe, these differences are also likely to affect signal detectability; how seriously, that depends on the specifics of the associated luminance transitions.

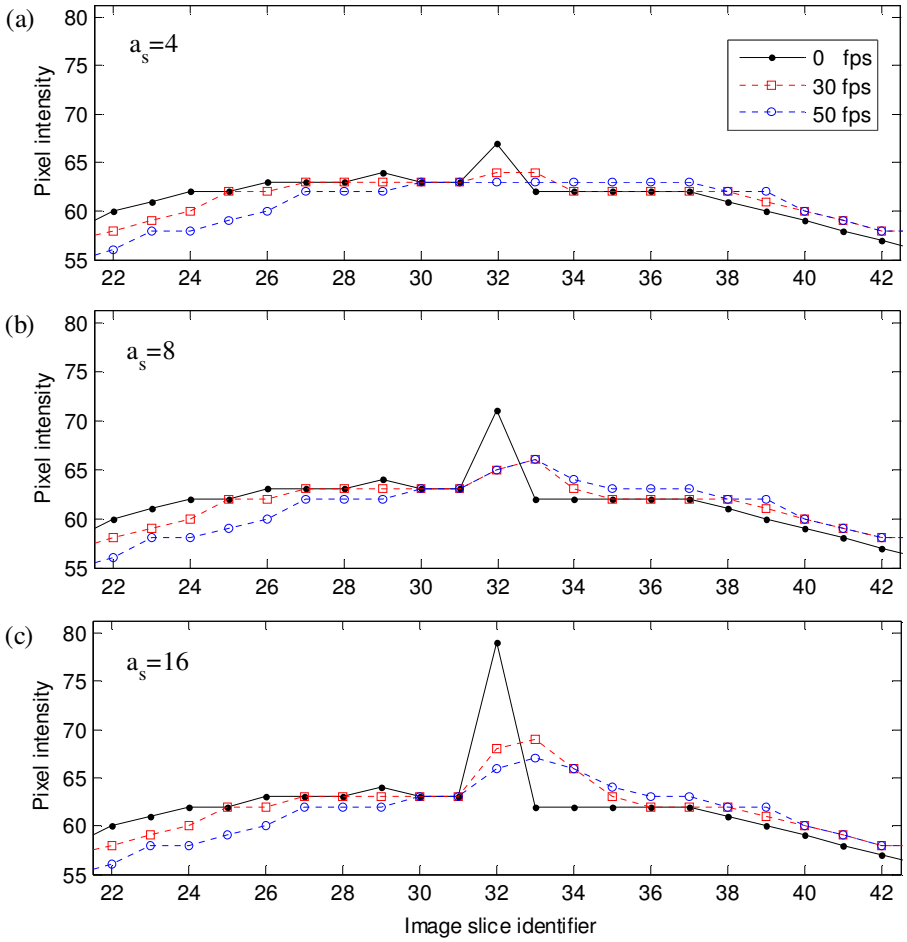


Figure 4.11: The intensity profile of the central pixel in a on-LCD image slice depicted across the sequence (from frame 22 to frame 42). In the corresponding pre-LCD image (which suffers no effects of the display), the 2D signal is present in the center of the central image slice. From top to bottom, the plots correspond to the three different contrasts of the signal: (a) $a_s = 4$, (b) $a_s = 8$, and (c) $a_s = 16$. For each signal contrast, the frame rate is varied from static (pre-LCD images), through 30 fps, and up to 50 fps.

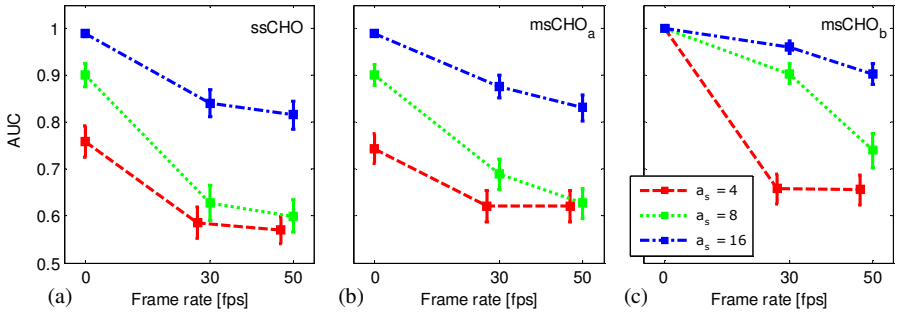


Figure 4.12: Model observer performance for the medical LCD characterized by the temporal response parameters from Figure 4.9: (a) single-slice CHO, ssCHO; (b) type a of the multi-slice CHO, $\mathrm{msCHO_a}$; (b) type b of the multi-slice CHO, $\mathrm{msCHO_b}$. For both msCHO models, the ROI is comprised of $R = 5$ consecutive slices centered around the slice in which the 2D signal is located. The a_s parameter describes the contrast of the signal (see text in Section 4.3.4.1 for details about the images).

4.3.5.2 Model observer performance

Observer performance results are shown in Figure 4.12. The three plots from (a) to (c) correspond to the three considered models: ssCHO, $\mathrm{msCHO_a}$ and $\mathrm{msCHO_b}$. Each plot depicts the model performance for three image data setups: pre-LCD, on-LCD-30 and on-LCD-50.

Overall, in line with the results of [Liang et al., 2008], all three models suffer a significant decrease in the detection performance as the frame rate is increased. This applies not only for the high frame rate of 50 fps but also for the medium rate of 30 fps. However, compared to the ssCHO, the msCHO designs, as expected, exhibit less sensitivity to the temporal effect of the display. Especially, when the contrast of the signal is reasonably high ($a_s = 8$ and $a_s = 16$), the performance of the $\mathrm{msCHO_b}$ seems least affected by the slow temporal response.

As we explain in Section 4.3.3 and more in detail in Section 3.3, the design approach of msCHO should allow a more accurate estimate of detectability compared to the ssCHO. The ssCHO estimate could be seen as somewhat pessimistic given that it has only restricted access to the image data (one slice only). With that assumption, we still find that, with reference to the pre-LCD images, the slow temporal response of the medical LCD could decrease the AUC detection performance of an observer by as much as 20% when the frame rate is 30 fps, or even more (by almost 30%) when the frame rate goes up to 50 fps. Certainly, these trends depend on the temporal characteristics of a given display as well as on the considered contrast level of the signal.

Importantly, we remark that in the study presented here the pixel luminance was assumed constant throughout the frame duration, equal to the value achieved at the

end-of-frame. Conversely, in practice, the luminance of a pixel changes gradually over a frame duration. Therefore, it might be more realistic to assume that the perceived luminance is the average of the begin-of-frame and the end-of-frame luminance, or even better, to actually compute the luminance at the end of each display refresh cycle T_{refresh}. In this sense, the results obtained here may be regarded as a somewhat optimistic estimate of the extent of detectability degradation caused by the slow temporal response of the medical display. In support of this notion, the study of [Liang et al., 2008] found less degradation in signal detectability when they assumed the frame luminance equal to the end-of-frame luminance compared to the case where they assumed it equal to the average luminance within a frame.

4.3.5.3 Size of the ROI

In our experiments, the number of slices used by the msCHO was varied among 3, 5 or 7 slices, and the slices were centered on the slice in which the signal is located. The preferred size of ROI was selected based on the observed AUC values such that further growing of ROI does not affect the observer performance significantly. For pre-LCD data the size of ROI was $R = 3$. For greater frame rates and lower contrast of the signal, the value of R slightly increased. In particular, for on-LCD-30 and for $a_s = \{4, 8\}$ the size of ROI increased to $R = 5$. When the frame rate was further increased to 50 fps, for the on-LCD-50 data set, the value of R remained the same for $\mathrm{msCHO_b}$ and it increased to $R = 7$ for $\mathrm{msCHO_a}$. In the case of the highest considered signal intensity, $a_s = 16$, the size of ROI is $R = 3$ for all considered frame rates.

In general, for both the $\mathrm{msCHO_a}$ and the $\mathrm{msCHO_b}$, the ROI used with the on-LCD image sequences is greater than the number of slices used with the pre-LCD images. This conforms to the discussion from Section 4.3.4.2 which explains the faint appearances of a one-slice signal over multiple neighboring slices caused by the display response time. For the msCHO, and likewise possibly for humans, this presence of the signal over multiple slices allows a higher confidence level of the observer and thus it may be seen as an aid to the detection process. Of course, what does not help the observer is the decreased contrast of the signal in on-LCD images. Moreover, though not directly the focus of the present study, we note that the "delayed" peak of the signal (the trailing effect) may affect the correct localization of the signal in on-LCD images.

In conclusion, based on our results here, we would have a slight preference for the $\mathrm{msCHO_b}$ over $\mathrm{msCHO_a}$ as the preferred model for future investigations around temporal response of LCD devices. The $\mathrm{msCHO_b}$ design seems able to differentiate between the frame rates also at lower contrast of the signal (*e.g.* for $a_s = 8$), and it requires fewer slices at high frame rates (*e.g.* at 50 fps).

4.3.5.4 Reflections on a related human observer study

Soon after our model observer study, the results became available for a related human observer study of [Badano, 2009]. There, human performance for a CRT medical display was compared to that of an LCD display, using the images very similar to our image data of $a_s = 8$ category (the models for image objects were the same, only peak signal intensity was $a_{\mathrm{human}} = 10$). In particular, a total of 13 human observers (imaging scientists and graduate students) read images at two frame rates, a medium rate of 20 fps and a fast rate of 50 fps. The experiments were conducted using a two-alternative forced choice (2AFC) procedure[14] and the performance was evaluated in terms of the proportion of the correct responses (PC, on the scale of 0 to 1). Under the 2AFC paradigm, the AUC equals the PC [Barrett and Myers, 2004].

For the CRT display, as expected, the mean difference in human performance at 20 fps and at 50 fps was very small (0.049). Therefore, we may consider it approximately free from the temporal effects. In our model observer study, this corresponds to the pre-LCD images ($f_{\mathrm{frame}} = 0$ fps).

Given the model observer results, also expected were the results for LCD where a markedly larger difference was found between the two frame rates. For the LCD, the mean difference in human performance at 20 fps and at 50 fps was 0.156. Given our model observer results from Figure 4.12 for the case of $a_s = 8$, the corresponding difference in AUCs between 30 fps and 50 fps is about 0.5 for $\mathrm{msCHO_a}$ and about 0.16 for $\mathrm{msCHO_b}$. Thus, with respect to differentiating between different non-zero frame rates, $\mathrm{msCHO_b}$ seems more similar to humans than $\mathrm{msCHO_a}$. This provides an additional argument for choosing $\mathrm{msCHO_b}$ over $\mathrm{msCHO_a}$ as the preferred model for observer experiments with on-LCD image data.

4.4 Preclinical validation of a novel LCD design

In the study described here, we evaluate the diagnostic performance of a novel LCD design targeted specifically at clinical viewing of a rather recent image acquisition modality: the three-dimensional (3D) digital breast tomosynthesis (DBT) images.[15] The design problem of interest concerns the slow response time of an LCD, previously noted in Section 4.3 as a factor of the diagnostic performance when browsing volumetric image datasets. We conduct a model observer study to compare two LCD designs: the LCD with the novel algorithm for automated compensation of the slow temporal response and the LCD without such compensation. The images in our study are real clinical tomosynthesis backgrounds (signal-absent images) with added simulated lesions (signal-present images). The effects of image displaying (LCD response

[14] In a 2AFC experiment, the observer is presented with two images, one of them is signal-present and another is signal-absent. The task for the observer is to identify which of the two is the signal-present image.

[15] Digital breast tomosynthesis has been approved by the U.S. Food and Drug Administration (FDA) on February 11, 2011.

time) are simulated following the process in Section 4.3.3. Compared to the study in Section 4.3, we now use a different model of LCD temporal response and different display parameters. Importantly, the results of our *preclinical* model observer study were used to pinpoint the characteristic frame rates for the subsequent *clinical* validation employing human observers in place of the models [Marchessoux et al., 2011]. We reflect on that human observer study in Section 4.4.4.

4.4.1 Study rationale

Today, 3D DBT is gradually taking over conventional 2D digital breast mammography (DBM) imaging. Admittedly, mammography has proved over years to be an effective imaging tool for detecting breast cancer at an early stage (breast cancer screening) [Tabár et al., 2011]. Nevertheless, as many as 20% to 30% of breast cancers remain undetected with mammograms [Rafferty et al., 2013].

One major drawback of the breast mammography is known as the "overlapping tissue" [Park et al., 2007a,Rafferty et al., 2013]. It is related to the density of breast tissue and how it is depicted on mammograms. Namely, the cause of the problem is in the process of acquisition of a breast mammogram. The breast is pressed between two flat plates so that the x-ray source is on one side (perpendicular to the compressed breast) and the detector is on the other. The critical factor is that the radiation source remains stationary during the imaging process and so only a single 2D projection image can be made, formed by absorption of x-rays at the detector.[16] In this way, structures of the radiographically dense tissue could be superimposed resulting either in an artifact which mimics an abnormality (causing a false-positive recall) or in masking the actual abnormality (causing a false-negative, *i.e.*, a missed cancer). By contrast, breast tomosynthesis allows the x-ray source to move around the breast and acquire a series of projection view (PV) mammograms which are then used to reconstruct the tomographic breast volume. Thereby, the effect of overlapping tissues is minimized which suggests potential for avoiding the aforementioned problems of mammography.

Therefore, the DBT can be expected to improve detection of the breast lesions, not only masses but also subtle microcalcification clusters. The results of [Andersson et al., 2008], for example, indicate that the cancer visibility on DBT is superior to DBM suggesting that tomosynthesis may have a higher sensitivity for breast cancer detection. Similarly, the simulation results of [Ma et al., 2008] show consistently that 3D DBT allows detection of smaller tumors and smaller microcalcifications than the 2D DBM images. Most recently, [Rafferty et al., 2013] found diagnostic accuracy for combined DBT and DBM to be superior to that of DBM alone. Moreover, in the screening setting, the addition of tomosynthesis significantly reduced recall rates for non-cancer cases.[17]

[16] Standard screening mammography includes two views (projections) of each breast: from above (cranial-caudal view, CC) and from an oblique or angled view (mediolateral-oblique, MLO). Diagnostic mammography may involve additional views.

[17] [Rafferty et al., 2013] note that the addition of tomosynthesis to the standard mammogram represents

With the current growing evidence of the practical diagnostic benefits of DBT, it is not surprising that this technology has drawn much attention from the medical imaging community, either in the domain of image acquisition and reconstruction, or in the field of image data presentation, and finally image interpretation. Our focus is on image presentation (medical display) and how it affects the interpretation of the images (detection of breast lesions). Commonly, the radiologists inspect the thin tomographic slice images while browsing an image sequence viewed on a liquid crystal displays (LCD) where slices of the reconstructed volume are shown sequentially, at an arbitrary frame rate. Importantly, despite the fact that the quality of LCDs has significantly improved over the last few years, the slow response times of liquid crystal cells remain a limiting factor for signal detection performance in the browsing mode of image reading at high frame rates [Liang and Badano, 2007, Liang et al., 2008]. As indicated by the human reader study by [Badano, 2009], the slow response of liquid crystal display devices reduces the detection performance when using high frame rates to inspect volumetric images in sequence-browsing presentation.

4.4.2 Experimental goal

We aim to evaluate the quality of a novel medical image display optimized for DBT (Barco MDMG 5221 display optimized for DBT). One of the major advancements of the new display device is the automated temporal response compensation. The novel algorithm is aimed to diminish the negative influence of the current LCD technology (slow temporal response) on signal detectability in the "fast" changing displayed image scenes, such as those in the browsing mode of image reading at high frame rates.

4.4.3 Study design and methodology

In order to assess the effects of the novel algorithm for LCD temporal response compensation, we compare the novel display to an existing state-of-the-art full field digital mammography (FFDM) display which has no such compensation. Hereafter, the two displays are referred to as the "motion compensated" LCD (mcLCD) and the "regular" LCD (regLCD), respectively.

Given that the main purpose of a medical display is to assist radiologists in diagnostic procedures, it appears most relevant to perform a task-based procedure for image quality assessment [Park et al., 2010]. Accordingly, we define the task of interest to be detection of breast lesions in tomosynthesis images viewed on the display under test.

For estimating the detection performance, we use the multi-slice channelized Hotelling observer of type b ($\mathrm{msCHO_b}$) defined in Section 3.3.3.3 of Chapter 3. This model design was suggested in Section 4.3 as the preferred one for the condition of image

additional radiation exposure to the patient. However, at this point in time, the standard mammogram is still necessary for comparison with prior examinations. Investigations are ongoing towards replacing the standard mammogram with a mammogram synthesized from the tomosynthesis images to reduce the dose.

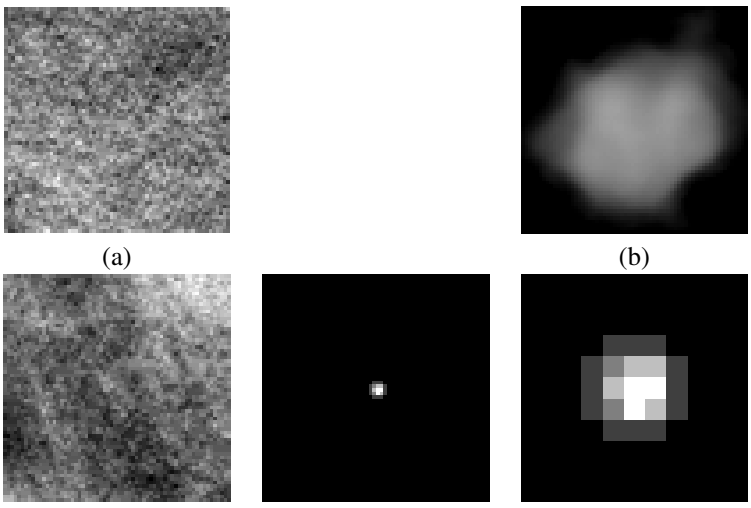


Figure 4.13: Image data used in the study (see text): (a) central slice of an example signal-absent image, (b) an example 2D lesion extracted from a real clinical DBM image, (c) central slice of an example signal-present image, (d) central slice of the volumetric mass synthesized from 2D lesion in (b), (e) enlarged mass area from (d).

browsing on a display with slow temporal response. The model observer experiments are conducted for clinical digital tomosynthesis images of the breast with added simulated mass lesions and simulated effects of the temporal response of the display. The details of the study are described next.

4.4.3.1 pre-LCD image data

The images described here, signal-present or signal-absent ones, are referred to as "static" or pre-LCD images as they do not take into account the temporal response of the display. We use a total of 6000 multi-slice images of $64 \times 64 \times 41$ pixel size, where $N = 41$ is the number of slices and $X = Y = 64$ denotes the width and height of each image slice. The background images are crops from reconstructed clinical digital breast tomosynthesis (DBT) images. The pixel values are coded in 10 bits.

Half of the backgrounds are used as signal-absent images, see an example in Figure 4.13 (a). To generate signal-present images, we add a synthesized volumetric mass (3D signal) in the central three slices of the remaining 3000 backgrounds. Figure 4.13 (c) shows the central slice of an example signal-present image from our experiments. The 3D signals are generated using the data set of 2D lesions extracted from real clinical digital mammography images, provided by Dr. Elizabeth Krupinski from The University of Arizona. First, a 2D lesion is warped using mathematical morphology operations to get a 3D shape. Then, the resulting volume is interpolated

between the slices in order to mimic the X-ray interaction (absorption). Finally, the lesion is smoothed to avoid any sharp gradient at the borders and it is normalized. An example 2D lesion and the central slice of its corresponding 3D lesion are depicted in Figure 4.13 (b), (d), and (e). The synthesized 3D mass breast lesion of a given density is inserted in the reconstructed background volume.[2]

4.4.3.2 LCD temporal response simulations

In contrast to the display model from Section 4.3.4.2 where we assumed each slice to be drawn on the screen only once for the frame duration (we considered only the pixel intensity values at the end of frame duration intervals), here we do a more truthful simulation by accounting for the fact that a given slice is (re-)drawn on the screen after every refresh time interval (we now consider the pixel intensity values at the end of each display refresh cycle).

In Figure 4.14 we show the measurement-based parameters of the investigated five-mega-pixel 10-bit grayscale medical LCD monitors: (left) the luminance response curve $c(g)$ and (right) the matrix of response times $\mathbf{Q}$. Note that in contrast to the measurements of [Liang and Badano, 2007] where the elements of $\mathbf{Q}$ were the percentage of the luminance target transition achieved after one frame duration, in this case the elements of $\mathbf{Q}$ are the actual response times of the corresponding transitions (times needed for the target transition to be completed). The values of response times are given in milliseconds (ms). Overall, by comparing matrix $\mathbf{Q}$ of the display in the previous study (see Figure 4.9 (b)) to the corresponding matrix of the displays here, we find the two rather similar in terms of the distribution of the response times across luminance (grayscale) transitions. The parameters in Figure 4.14 refer to the LCD with no compensation for the slow temporal response.

Two points are obvious from the response times in matrix $\mathbf{Q}$. First, the "rise"-times (transitions from lower to higher intensity levels) tend to take longer than the "fall"-times (transitions from higher to lower intensity levels). Overall, the transitions from very low to medium-high intensities take the longest (even up to 50 ms) while the shortest are transitions from very high to very low luminance levels (white to black grayscale transition). Second, the majority of transitions takes longer than the time between the two consecutive display refresh cycles $T_{\mathrm{refresh}} = 20$ ms (corresponding to $f_{\mathrm{refresh}} = 50$ Hz). This observation clearly suggests that, depending on the frame rate, it is possible that the target luminance level $l^0(n+1)$ could not be achieved during the frame duration, T_{frame}. That is to say, when the frame rate is "high" such that the frame duration $T_{\mathrm{frame}} = {}^1/_{f_{\mathrm{frame}}}$ is shorter than the response time for the particular luminance transition $\triangle l(n+1, n) = l^0(n+1) - l(n, T_{\mathrm{frame}})$, the actual displayed luminance of the pixel will be different (lower or higher, depending on the sign of $\triangle l$) from the target luminance level $l^0(n+1)$. As will be shown in our experimental

[2] Subsequent to our study, the team including some of our collaborators here [Vaz et al., 2011] has developed an improved method for generation of the 3D mass breast lesions where the lesion is inserted directly in the projection images.

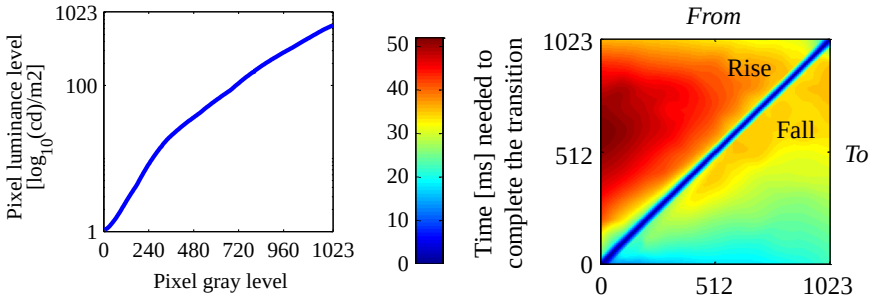


Figure 4.14: Parameters of the LCD monitors in the study. Left: Luminance response curve of the display, $c(g)$. Right: Matrix of the liquid crystal director reorientation times, **Q** [ms]. The elements of the matrix represent the time needed to achieve the complete transition *from* the luminance level determined by the x-coordinate of the matrix, $l_{\text{from}} = c(g_x)$, *to* the luminance level determined by the y-coordinate of the matrix, $l_{\text{to}} = c(g_y)$. The label "Fall" in the bottom-right triangle of the matrix suggests the transitions from higher to lower intensity levels, and the label "Rise" in the top-left triangle of the matrix suggests the transitions from lower to higher intensity levels. The values in the matrix correspond to the LCD display with no compensation for the slow temporal response.[18]

results in Section 4.4.4, this consequence of the slow LCD response time can play an important role in the signal detection performance, especially at higher frame rates.

Next, we explain the details of the temporal response model of [Wang et al., 2004a]. According to the description in Section 4.3.3, the model is used to estimate the actual achieved display pixel intensity at a given point in time. [Wang et al., 2004a] make use of the small angle approximation to derive the analytical relationship between the liquid crystal director reorientation time and its consequent optical rise and decay (fall) times. As the authors show, if a liquid crystal element is initially biased at voltage V_1, and the voltage is removed instantaneously at the time instance $t = 0$, the transient phase change δ at the time instance t, can be approximated as

$$\delta(t) \cong \delta^0 \exp\left(-\frac{2t}{\tau^0}\right). \tag{4.8}$$

Here δ^0 is the phase change corresponding to the complete transition from V_1 to the zero voltage V_0; assuming a 10-bit display and using $|\cdot|$ to denote the absolute value, it can be computed as $\delta^0 = (\pi/2^{10})\,|\triangle l(n+1, n)|$. The value of τ^0 is the time needed for the LCD cell to complete the transition from V_1 to V_0. Given our display measurements in Figure 4.14, τ^0 takes its value from the matrix **Q**, *i.e.*, $\tau^0 = q\left(g(n, T_{\text{frame}}), g^0(n+1)\right)$. According to [Wang et al., 2004a], the change in intensity of a switching LCD cell which corresponds to the phase change δ can be

Table 4.3: Parameters of the display simulations. We assume native display parameters of $f_{\mathrm{refresh}} = 50$ Hz and $T_{\mathrm{refresh}} = 20$ ms.

	Database name	Frame rate f_{frame} [fps]	Frame repeat FR	frame duration T_{frame}
0	pre-LCD	-	-	-
1	on-LCD-10	10	5	$5 \times T_{\mathrm{refresh}}$
2	on-LCD-13	12.5	4	$4 \times T_{\mathrm{refresh}}$
3	on-LCD-17	16.67	3	$3 \times T_{\mathrm{refresh}}$
4	on-LCD-25	25	2	$2 \times T_{\mathrm{refresh}}$
5	on-LCD-50	50	1	$1 \times T_{\mathrm{refresh}}$

modelled as:

$$I(t) = \sin^2\left(\frac{\delta(t)}{2}\right). \tag{4.9}$$

Thus, the index of completeness of the target luminance transition (see in Section 4.3.3) is the ratio of the intensity change achieved at time instance t over the target intensity change,

$$\rho(t, q(n+1, n)) = \frac{\sin^2\left(\frac{\delta(t)}{2}\right)}{\sin^2\left(\frac{\delta^0}{2}\right)}. \tag{4.10}$$

This describes the temporal response model of the regLCD device.

For mcLCD, in order to reduce the temporal effect, an overdriving value within one frame is introduced [Kimpe and Marchessoux, 2010]. In that way, the target values are reached with a special processing and any enhancement of the temporal noise is avoided. The solution for temporal response improvement does not introduce any artifacts by avoiding any overshooting. The exact details of the motion compensation algorithm are the subject of a patent application [Kimpe and Marchessoux, 2010].

In our study, the regLCD and mcLCD are compared for five different frame rates in the range of 10 to 50 fps. The corresponding sets of images (after display simulation) are referred to as image databases and denoted on-LCD-10, for f_{frame}=10 fps (frame repeat $\mathrm{FR} = 5$), through on-LCD-50, for f_{frame}=50 fps ($\mathrm{FR} = 1$). We assume the native LCD parameters of $f_{\mathrm{refresh}} = 50$ Hz and $T_{\mathrm{refresh}} = 20$ ms. The details about temporal response simulation parameters are summarized in Table 4.3. Each database is created for both display models: the with- and the without motion compensation LCD. In addition, as a point of reference, we consider the "static" (pre-LCD) images where no display effects are considered. Thus, in total, there are $5 \times 2 + 1 = 11$ image databases in our experiments.

4.4.3.3 Observer performance experiments

We evaluate the quality of the displays based on the criterion of signal detectability, using the $\mathrm{msCHO_b}$ model defined in Section 3.3.3.3 and suggested in our previous study in Section 4.3 as the preferred model for studies of slow temporal response of medical LCD monitors.

The signal in our pre-LCD images only exists in the central image slice and thus fewer slices are considered for the ROI. In particular, the number of slices in the ROI is varied among 3, 5 and 7 slices adjacent to the signal slice (slice 21 out of a total of $N = 41$ slices). The channels used in the study are the first $P = 10$ dense difference-of-Gaussian (DDOG) channels. The values of our DDOG channel parameters correspond to those used in the study of [Abbey and Barrett, 2001] which showed to closely track human observer performance.

In our experiments, for on-LCD image data, the $\mathrm{msCHO_b}$ performance is computed for the pixel values achieved at the end of each refresh cycle during the T_{frame}. For example, when the frame repeat $\mathrm{FR} = 3$ (see Table 4.3), the detection performance is computed for on-LCD image values at the end of each $1 \times T_{\mathrm{refresh}}$, $2 \times T_{\mathrm{refresh}}$ and $3 \times T_{\mathrm{refresh}}$.

The experiments are MRMC studies with $N_{\mathrm{rd}} = 5$ readers per image database. Remember from Section 4.4.3.2 that a total of 11 different image databases is considered in the study: the pre-LCD database and five sets of on-LCD images for each mcLCD and regLCD(on-LCD-10, on-LCD-13, on-LCD-17, on-LCD-25, on-LCD-50). Each reader is trained on an separate subset of $N_{\mathrm{tr}} = 500$ trainer image pairs and applied on a unique set of $N_{\mathrm{ts}} = 500$ tester image pairs from a given database. The trainer and the tester images do not overlap. As a figure of merit for our MRMC experiments, we use the AUC in combination with the one-shot method [Gallas, 2006] for variance analysis.

4.4.4 Results and discussion

Similar as with the previous study, we examine the following aspects of our experimental data: (1) how the on-LCD data differs from its pre-LCD input depending on the details of the LCD temporal response, (2) how the signal detectability changes with increasing speed of sequence-browsing, as well as (3) how the msCHO performance compares to the human performance measured in a related human observer study from the literature [Marchessoux et al., 2011].

4.4.4.1 Simulated on-LCD image data

First, we performed simulations of the LCD effects in order to generate the on-LCD images for the model observer study. Figure 4.15 depicts the results of those simulations for the two displays in the study: the one with the motion compensation, mcLCD, and the other without such compensation, regLCD. The plots show the changes in

intensity of the central pixel in the xy-plane of an example signal-present image sequence. The five plots correspond to the five different frame rates in the study ranging from a relatively low 10 fps (top plot) up to the maximum achievable 50 fps (bottom plot). Overall, we observe that the on-LCD pixel intensity profiles of the two displays are different, less for the frame rates of 10 fps and 12.5 fps and more so for the high rates of 25 fps and especially 50 fps. This is due to the fact that for the lower frame rates (top plots) the same slice is refreshed more often than for the high frame rates (bottom plots) and so there is more time available for the LCD cells to attempt reaching to the target luminance level. Remember Eq. (4.3) and the point from Section 4.4.3.2 that each slice is displayed (refreshed) $\mathrm{FR} = f_{\mathrm{refresh}}/f_{\mathrm{frame}}$ times, where in our case $f_{\mathrm{refresh}} = 50$ Hz.

4.4.4.2 Model observer performance

Next, we perform a model observer MRMC study in order to evaluate the effects of the slow temporal response of regLCD and examine the potential benefit of motion compensation in mcLCD. Here, the model observer performance is computed after each refresh cycle within a given frame duration, *e.g.*, after each $kT_{\mathrm{refresh}} \leq T_{\mathrm{frame}}$ interval, where $k = 1, ..., T_{\mathrm{frame}}/T_{\mathrm{refresh}}$. For instance, when $T_{\mathrm{frame}} = T_{\mathrm{refresh}}$ corresponding to $\mathrm{FR} = 1$, we only compute the $\mathrm{msCHO_b}$ performance for on-LCD frames after the first refresh cycle, $T_{\mathrm{refresh}} = T_{\mathrm{frame}}$. For higher values of T_{frame}, *e.g.*, $T_{\mathrm{frame}} = 3T_{\mathrm{refresh}}$ or $\mathrm{FR} = 3$, the $\mathrm{msCHO_b}$ performance is computed for on-LCD frames at each $1T_{\mathrm{refresh}}$, $2T_{\mathrm{refresh}}$ and $3T_{\mathrm{refresh}}$. The results of these computations for each of the five on-LCD databases (on-LCD-10, on-LCD-13, on-LCD-17, on-LCD-25, on-LCD-50) together with those for the pre-LCD images, all for both regLCD and mcLCD, are presented in Figure 4.16. Note again that the value of FR is related to the frame rate by Eq. (4.3); for example, the value of $\mathrm{FR} = 2$ corresponds to $f_{\mathrm{frame}} = {}^{50}/_{2} = 25$ fps (*i.e.* the database on-LCD-25). The AUC values and their corresponding error bars depicted in the plot from Figure 4.16 are estimated using the one-shot algorithm [Gallas, 2006]. Additionally, for on-LCD images, we show the mean values of AUC scores computed at the end of each $T_{\mathrm{frame}}/T_{\mathrm{refresh}}$ refresh cycles.

Overall, we observe a clear drop in the AUC values for on-LCD images compared to those for the pre-LCD data, from AUC≈ 0.87 for pre-LCD images (DB-static) down to AUC< 0.83 in any of the on-LCD images. Moreover, the AUC trends suggest degradation in the detection performance of the observer as the frame rate is increased from 10 fps ($\mathrm{FR} = 5$) to 50 fps ($\mathrm{FR} = 1$). These findings are in line with the model observer predictions of [Liang et al., 2008] and those from Section 4.3 of this thesis, as well as with the human scores from the study of [Badano, 2009]. Importantly, we observe that the $\mathrm{msCHO_b}$ performance decreases less for the LCD with temporal response compensation, mcLCD, than for the LCD with no temporal response compensation, regLCD; see especially the mean AUC estimates marked by stars in Figure 4.16. This suggests that the display with temporal response compensa-

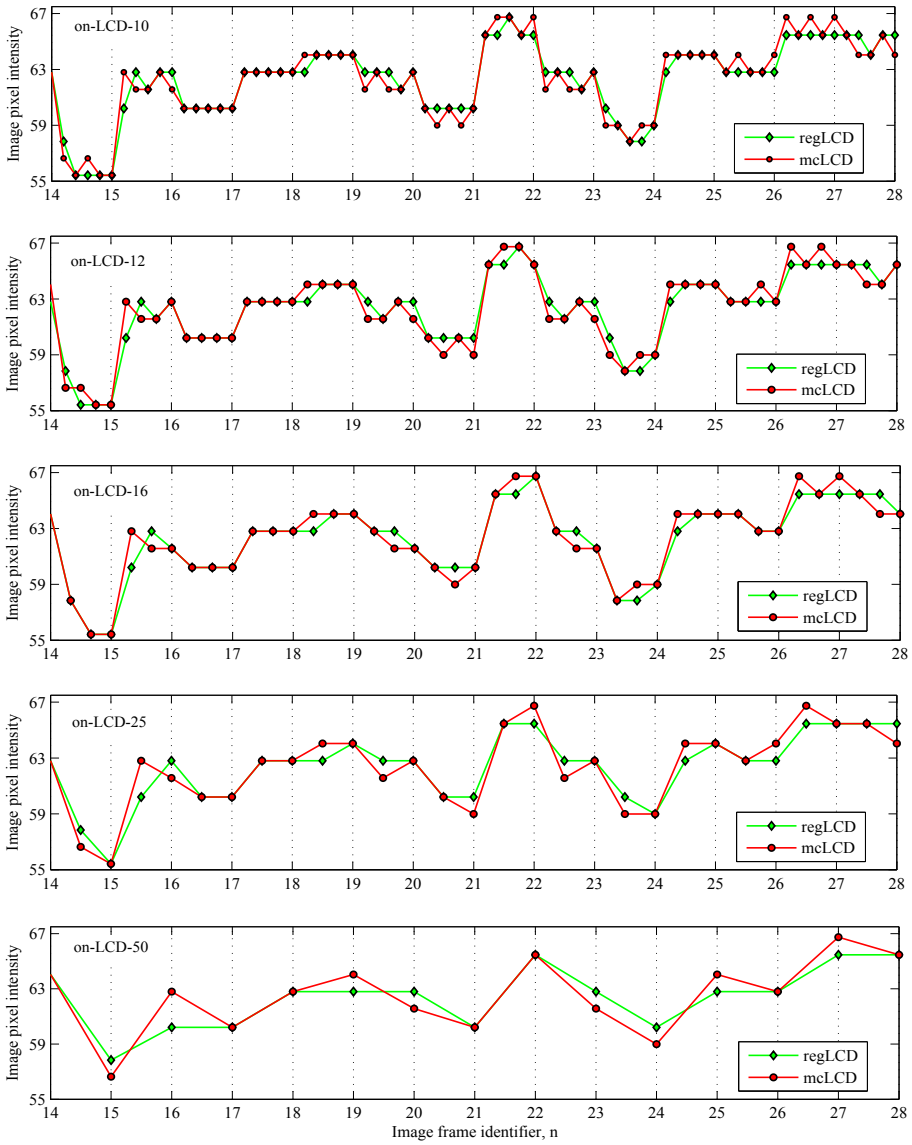


Figure 4.15: Pixel luminance values in on-LCD images, mcLCD compared to regLCD. For one signal-present image used in the study, the intensity profile of the central pixel (xy-plane) across slices 14 through 27 (z-direction) is shown. In the corresponding pre-LCD image, the signal is centered in the slice 21. Five different frame rates are considered (top to bottom): 10 fps, 12.5 fps, 16.67 fps, 25 fps, and 50 fps (see also Table 4.3). The values of image frame identifier (x-axis labels) denote the start of the frame duration for a given frame.

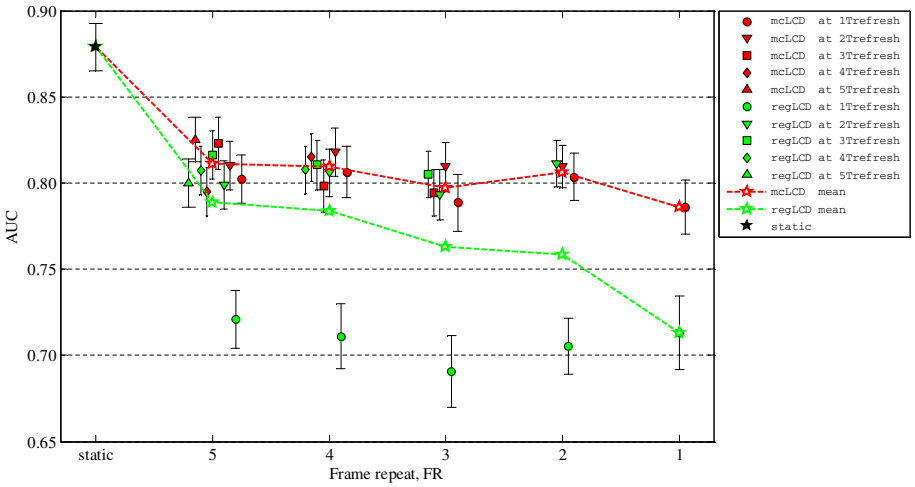


Figure 4.16: Detection performance of the $\mathrm{msCHO_b}$ for two displays: mcLCD, with temporal response compensation (DBT display) and regLCD, with no temporal response compensation (FFDM display). The $\mathrm{msCHO_b}$ performance is computed after each $kT_{\mathrm{refresh}} \leq T_{\mathrm{frame}}$ interval, $k = 1, ..., T_{\mathrm{frame}}/T_{\mathrm{refresh}}$ (see text for details). The computations are performed in an MRMC study with $N_{rd} = 5$ readers, each trained with a separate subset of $N_{tr} = 500$ training image pairs and all reading the same test set of $N_{ts} = 500$ test image pairs. The size of ROI used in $\mathrm{msCHO_b}$ computations is $R = 5$. The error bars are ± 2 standard deviations estimated by the one-shot method [Gallas, 2006].

tion could allow higher detectability of lesions and hence higher diagnostic accuracy in the sequence-browsing mode of image reading.

Finally, based on our model observer results, we select the frame rates of interest for the subsequent clinical validation study with the human observers [Marchessoux et al., 2011]. There, we choose the two frame rates for which the difference in AUC between mcLCD and regLCD is the largest: 25 fps ($\mathrm{FR} = 2$) and 50 fps ($\mathrm{FR} = 1$). Knowing that humans often achieve lower AUC values than the CHO models, we omit the frame rates at which the model captured smaller differences in performance of the two displays and go for the rates which could be expected to make a more obvious demonstration of the effect.

4.4.4.3 Reflections on a related human observer study

In line with our model observer study, the human observer study reported in [Marchessoux et al., 2011] was able to demonstrate the clinical benefit of reading the DBT images on the display with the motion compensation over reading them on the regular mammography display. The slices were shown in dynamic cine loops, at the two frame rates suggested by our model observer experiments. Specifically, the ben-

efit of mcLCD over regLCD was shown by as much as the 10% improvement in the radiologists' performance at a frame rate of 25 fps (the difference in AUCs was statistically significant), and by the 6% improvement at 50 fps (though not statistically significant).[19]

Knowing that experiments with medical experts often take a significant amount of both time and money, it is evident that the savings from using models to narrow down the scope of human experiments is most valuable. For example, in the case of our study, the number of test parameters for the human study was reduced by a factor of three compared to the study with the models (going down from six to only two values). Consequently, the required radiologist time was three times less than it would have been without the preceding model observer study. This reduction in the scope of experiments brings multiple benefits. On the practical side, needing less of radiologists' time reduces expenses and also requires less time to complete the experiment. In addition, asking the radiologists for less of their time makes them more readily available which shortens the overall time line of the study. Moreover, from the psychological point of view – which must not be neglected in this kind of experiments, humans in general tend to be more willing and (able to stay) involved in the experiments which take a shorter rather than a longer time.

4.5 Upsampled msCHO design for LCDs with slow temporal response

In this section, we extend the current implementation of the msCHO model to incorporate within-frame luminance information (see Figure 4.7 for illustration). We refer to the new model as the *upsampled* msCHO, umsCHO. The two approaches, msCHO and umsCHO, are compared on a set of synthesized 3D images under the frame rates of 16.67, 25 and 50 fps. In order to investigate the influence of the luminance change profiles on the performance of the two models, we consider two different temporal response models of an LCD: a linear model by [Liang et al., 2008] and the model proposed by [Wang et al., 2004a]. The two models have been introduced in Section 4.2 and Section 4.4, respectively. In addition, as a point of reference in evaluating the performance of the two models, we consider the case in which the browsing effects are ignored ("static" display mode).

[19] According to [Marchessoux et al., 2011], this low statistical significance was also correlated with the feedback from the observers that 50 fps was too fast for the detection task. The numerical results confirmed that the observers had a tendency to give lower confidence score when complex backgrounds were viewed at the frame rate of 50 fps. The authors also remarked that the radiologists in the study were trained mammographers who were not well accustomed to image interpretation by browsing image sequences, as it is done in digital breast tomosynthesis (DBT) but not in digital breast mammography (DBM). At the start of the study (year 2010), none of them had extensive DBT training due to the novelty of the modality.

4.5.1 Study rationale

As explained in Section 4.3.3 and illustrated in Figure 4.7, terms of *within-frame* and *end-of-frame* luminance values are used to refer to the luminance achieved, respectively, at the end of each frame duration, T_{frame} and at the end of each display refresh cycle T_{refresh} which occurs during T_{frame} duration, *i.e.*, at time instances for which $t = kT_{\mathrm{refresh}} < T_{\mathrm{frame}}$, $k \in \mathcal{N}$. In the example from Figure 4.7, the end-of-frame luminance values are measured at the time instances corresponding to each $T_{\mathrm{frame}} = 3T_{\mathrm{refresh}}$ while the within-frame luminance values correspond to the time instances of T_{refresh} and $2T_{\mathrm{refresh}}$.

In Section 4.3 and Section 4.4, the msCHO was restricted by design to the analysis of the luminance of a display pixel at the end of the frame duration (end-of-frame luminance) while ignoring the information about the luminance transition over the frame duration (within-frame luminance). Only the end-of-frame luminance was considered also in the study in Section 4.3 as well as in the earlier study by [Liang et al., 2008]. Nevertheless, in the latter two studies this restriction came from the display model (it concerns only the end-of-frame luminance values) rather than from the msCHO directly.

One weakness of the methods which ignore the within-frame luminance is their inability to differentiate between, for example, two displays with different profiles of luminance over time as long as their end-of-frame luminance levels are the same. Moreover, such methods are inadequate to capture the full effects of the techniques for response time compensation (overdrive technologies) used in today's high-end medical LCDs [McCartney, 2003, Kumar et al., 2005]. At the same time, studies with humans indicate a clear benefit of applying such techniques [Marchessoux et al., 2011].

4.5.2 Experimental goal

The objectives of the present study are twofold: (1) to assess the significance of incorporating the within-frame information when estimating the detection performance in a sequence-browsing image reading scenario, and (2) to examine the role of the luminance transition form (profile) of an LCD on the estimated performance.

4.5.3 Novel observer model: upsampled msCHO (umsCHO)

So far, in Section 4.2 and Section 4.4, our study of the effects of slow LCDs on signal detectability using the msCHO strategy was restricted to the analysis of the *end-of-frame* luminance values $l_{\mathrm{end}} = l(mT_{\mathrm{frame}})$, $m = 1, 2, \ldots$. Now, we incorporate additional data given by the *within-frame* information, $l_{\mathrm{in}} = l(nT_{\mathrm{refresh}})$, $n = 1, 2, \ldots$ and $nT_{\mathrm{refresh}} \neq mT_{\mathrm{frame}}$. For this purpose, we modify the msCHO model design to process image values after each T_{refresh} interval, as illustrated in Figure 4.17. The new model is named the *upsampled* msCHO (umsCHO).

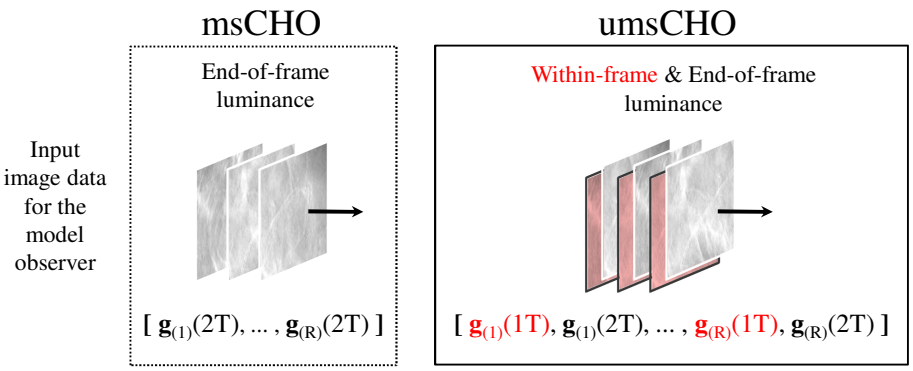


Figure 4.17: Upsampled msCHO (umsCHO) model design adapted from the msCHO model introduced in Section 3.3.3. While the msCHO is limited to the end-of-frame on-LCD image data, the umsCHO model has access to both the within-frame and the end-of-frame on-LCD image data. Specifically, the illustration assumes a frame rate of $f_{\text{frame}} = 25$ fps under the display refresh rate of $f_{\text{refresh}} = 50$ Hz; hence each image frame (slice) is displayed over the frame duration $T_{\text{frame}} = 2T_{\text{refresh}} = 2T$. There, the msCHO acts only on image pixel values at the end of frame duration intervals ($2T$, $4T$, ...), while the umsCHO is aware of the image values at each refresh time interval T.

Compared to the msCHO, the new umsCHO model has access to the image information sampled over more finely spaced intervals of time and thus we expect it to make more accurate estimates of the detectability in on-LCD images. Moreover, the conditions observed in human trials, where the luminance is not a discrete but a continuous function of time, are better approximated by the umsCHO sampling the time domain more frequently than the msCHO. The two models are explored in more detail in our experiments described next.

4.5.4 Study design and methodology

In the following, we describe (1) the parameters of pre-LCD image data used in the experiments, (2) the process of LCD display simulations, and (3) the details of MRMC model observer experiments.

4.5.4.1 pre-LCD image data

The pre-LCD images correspond to those in Section 4.3.4.1. There is a total of 2200 image samples of which 1100 signal-absent and 1100 signal-present ones. The volumes are $256 \times 256 \times 64$ pixels in size, where $N = 64$ is the number of slices. The background images are synthesized as 3D CLB [Bochud et al., 1999] with a 2D designer nodule [Burgess et al., 2001, Liang et al., 2008] added to the central slice of the

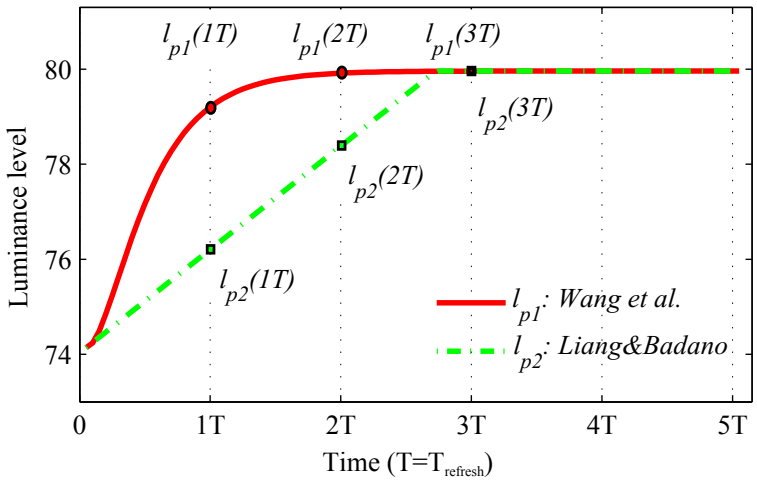


Figure 4.18: Pixel luminance change at different on-LCD frames. Two luminance transition models are depicted: l_{p1} corresponding to the work of [Wang et al., 2004a] and l_{p2} corresponding to the work of [Liang et al., 2008].

signal-present sequences. The details can be found in Section 4.3.4.1.

4.5.4.2 LCD temporal response simulations

As in the previous two studies in Section 4.3 and Section 4.4, we restrict our analysis of the display's effects to the response time of liquid crystal cells, the temporal response of the LCD. In all other aspects including spatial noise and contrast variability due to viewing angle, the display performance is considered ideal. Thus, the difference between our pre-LCD image and its corresponding on-LCD image is in the luminance of their pixels (refer to Figure 4.7 for illustration). The display parameters in our display simulations correspond to a 5MP 10-bit medical grayscale LCD for full-field digital mammography from the study in Section 4.4. Accordingly, the luminance response curve $c(g)$ and matrix $\mathbf{Q}$ of the liquid crystal director reorientation times correspond to those in Figure 4.14.

In this study we explore two different models for the temporal response of an LCD (see Figure 4.18): l_{p1}, a physics-based profile model proposed by [Wang et al., 2004a], and l_{p2}, a linear profile model proposed by [Liang et al., 2008]. We note from Figure 4.18 that the level of luminance achieved at the end of each display refresh interval largely depends on the form of the luminance transition curve: $(1 - \exp(-2t/\tau^0)$ in the case of l_{p1}, or t/τ^0 for l_{p2}, where τ^0 denotes the reorientation time of the liquid crystal director. Here, the values of parameter τ^0 are determined using measured values from the matrix of the response time shown in the right of Figure 4.14.

Thus, to summarize, we start by generating a pre-LCD image in grayscale space,

as explained in Section 4.5.4.1. Next, the pre-LCD image pixel values are converted to luminance values (target luminance images) using the luminance response curve from the left of Figure 4.14. Then, we apply the two temporal response models of medical LCDs described in the previous two sections to obtain two on-LCD images in luminance space (achieved luminance images). Specifically, three different frame rates $f_{\mathrm{frame}} = \{16.67, 25, 50\}$ fps are simulated with each LCD model. The parameters of the LCD temporal response correspond to those in the right of Figure 4.14. In this way, we create a total of seven image data sets – the pre-LCD ("static") and six on-LCD sets, corresponding to three frame rates $f_{\mathrm{frame}} = \{16.67, 25, 50\}$ fps for each l_{p1} and l_{p2}.

4.5.4.3 Observer performance experiments

For each model msCHO and umsCHO, we conduct the experiments for the seven aforementioned image setups, one pre-LCD and six on-LCD parameter configurations. All experiments are MRMC studies with $N_{\mathrm{rd}} = 5$ readers per image category, each trained on a separate subset of $N_{\mathrm{tr}} = 200$ trainer image pairs and applied on the set of $N_{\mathrm{ts}} = 100$ tester image pairs. The training and the testing images do not overlap and all readers read exactly the same set of tester images. The observer performances are compared in terms of the detection SNR computed from the observer's test statistics, using the Eq. (3.15). The corresponding error bars are estimated using an MRMC-type of bootstrap analysis [Beiden et al., 2000, Gallas et al., 2009] where each bootstrap iteration selects a set of readers and cases.

4.5.5 Results and discussion

Our data analysis addresses two effects of the LCD temporal response: (1) the effect on image data and (2) the effect on model observer performance.

4.5.5.1 Simulated on-LCD image data

Figure 4.19 shows the results of our display simulations following the two temporal response models from Figure 4.18. The plots present the changes of the intensity of the central pixel in the xy-plane for one example signal-present image sequence. Each plot corresponds to three different frame rates, $f_{\mathrm{frame}} = \{16.67, 25, 50\}$ fps. Next to the changes in intensity of the on-LCD images, also the intensity profile of the corresponding pre-LCD image sequence is shown, to serve as a reference for comparative analysis. Two main observations can be made from these plots: (1) the profile of the luminance transitions over time has an impact on the within-frame pixel values achieved while browsing through an image sequence, and (2) as the frame rate increases (from 16.67 to 50 fps), the error in the estimated on-LCD image values introduced by ignoring the within-frame image values increases. For the specific display parameters and the two luminance transition models l_{p1} and l_{p2} in our study, the difference between the pixel intensity profiles for the l_{p1} and l_{p2} appear quite significant.

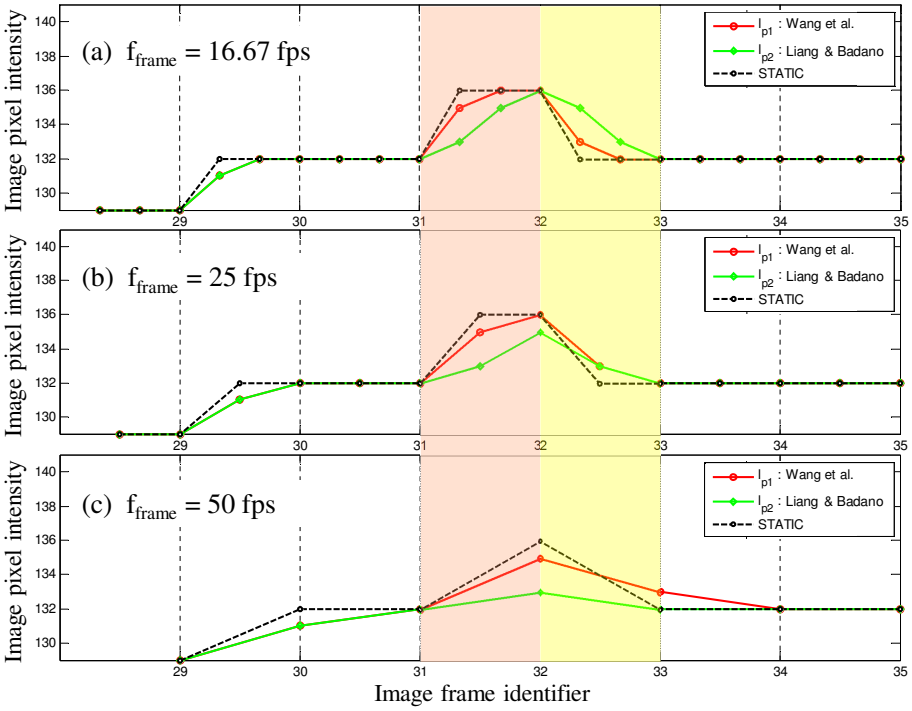


Figure 4.19: The intensity profile of the central pixel in the image slice is depicted across one image sequence. The frame rate is varied from (a) $f_{\text{frame}} = 16.67$ fps, through (b) $f_{\text{frame}} = 25$ fps and up to (c) $f_{\text{frame}} = 50$ fps. For each frame rate, two luminance transition models are considered, l_{p1} and l_{p2} (see Figure 4.18). As a reference, the dashed line in each plot represents the intensity profile of the static (pre-LCD) image. Compared to their pre-LCD values, on-LCD images exhibit a trailing effect in the slices around the signal. This is caused by slow temporal response of the display.

Therefore, we may expect also the associated levels of detectability to be different. This is investigated next.

4.5.5.2 Model observer performance

In Figure 4.20 we show the results of the msCHO and umsCHO experiments when the size of ROI is 7. The top plot depicts the results for l_{p1} model of luminance transitions proposed by [Wang et al., 2004a], and the bottom plot gives the results for l_{p2} linear luminance transition model. Indicated error bars correspond to ± 2 standard deviations estimated using bootstrap with 1000 re-samplings. Overall, the SNR trends suggest a degradation in detection performance as the frame rate increases, which is consistent with the results of our studies in the previous two sections as well as with the related

study in the literature [Liang et al., 2008].

However, given the luminance change profiles from Figure 4.18 and the changes in intensity of the central signal pixel illustrated in Figure 4.19, we expect the detectability to decrease as the frame rate increases. This expectation is confirmed by umsCHO but not by msCHO. For our experimental data (see Figure 4.19), the difference between msCHO and umsCHO input data is the greatest for the frame rate of 25 fps. This is explained by the fact that the msCHO only knows the nearly asymptotic luminance values achieved at 2T, while the umsCHO is also aware of the much lower values achieved by 1T. This causes the msCHO to overestimate the detection performance at 25 fps.

Finally, by comparing the SNR performance for the two luminance profiles, l_{p1} and l_{p2} (see the top versus the bottom plot in Figure 4.20), we notice that it is lower for the linear l_{p2} profile, for both msCHO or umsCHO. Especially, at the higher frame rates of 25 and 50 fps, the difference between SNR values for l_{p1} and for l_{p2} is approximately 2. This clearly suggests that an adequate choice of the luminance model in simulations of the effects of the LCD luminance temporal transitions is essential for a reliable estimate of the effects of slow LCD response time on the detection performance in the sequence-browsing mode of image viewing.

Given that the luminance is a continuous function of time, the umsCHO could be a preferred human-like model design over the msCHO as it samples the time domain more frequently. While further investigation is needed to validate the agreement between the performance of the proposed umsCHO model and that of a human, our results indicate promise for the methodology to be used with clinically relevant image data as a measure of detection-based image quality.

4.6 Single-slice versus multi-slice image viewing

We conduct a series of human observer experiments in order to assemble data about human performance for different levels of task difficulty. We compare single-slice versus multi-slice sequence-browsing mode of image viewing. The results are intended to guide future work towards designing a human-like model observer for volumetric image data. As outlined in Chapter 3, one possible approach to designing such a model would be to modify some of the existing models, *e.g.*, the msCHO models from Section 3.3.3, such that they can better predict the human performance. However, before we would modify the models, it is necessary to first identify the key factors of their required behavior.

4.6.1 Study rationale

As remarked on several occasions throughout this thesis, modelling humans requires first collecting human data. In the previous studies in this chapter, we ran experiments with the model observers and reflected on related human data when that was possible.

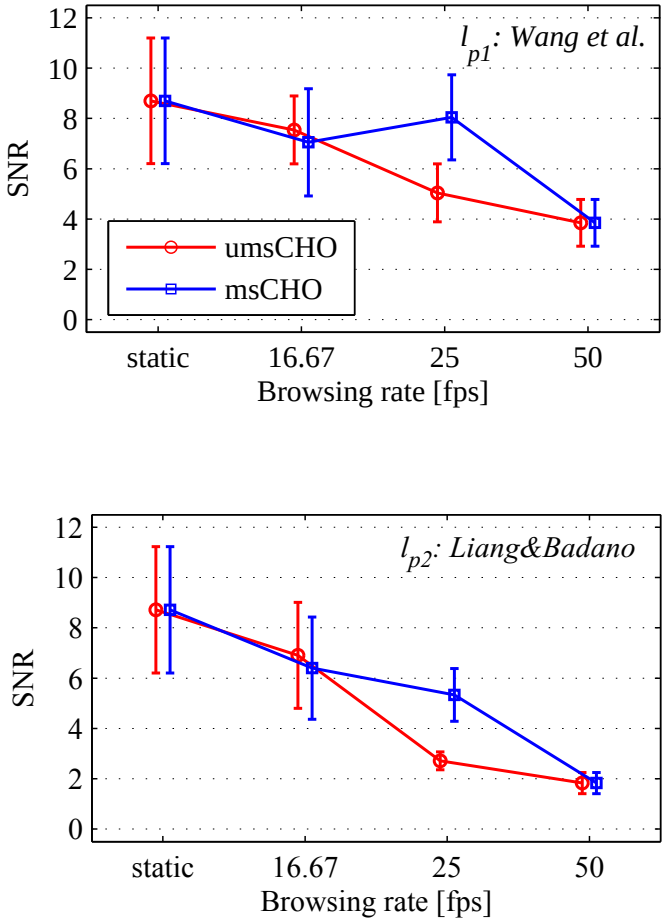


Figure 4.20: Detection performance of the umsCHO compared to the msCHO for two different luminance transition profiles: (top) l_{p1} corresponding to [Wang et al., 2004a] and (bottom) l_{p2} corresponding to [Liang and Badano, 2007].

Encouragingly, we found the models to agree with humans reasonably well, at least in terms of how they rank different image configurations, *i.e.*, the corresponding system parameters. It is of particular interest to find such an agreement for the volumetric image data – there is little evidence of similar studies in the literature to date.

Nevertheless, concerning the detection of volumetric signals, the human studies which we refer to as well as the several other studies in the literature are typically targeted at evaluation of the specific images, or systems. These results, despite undoubtedly very useful, could be overlooking some important aspects of human performance in the scenario of sequence-browsing. We believe that, on the way to designing a model (or a class of models) which could reliably predict human performance in the task of volumetric signal detection (in the browsing mode of image viewing), it is necessary to first understand some of the main factors which could be influencing that performance.

The study reported here as our contribution to uncovering the mechanisms behind one such factor – the *difficulty* of the detection task. Not unexpectedly, our results in Chapter 3 indicate that the difference between 2D and 3D task performance is affected by the properties of the image data which in fact determine the "difficulty" of the task. Specifically, the disparity between 2D and 3D task performance goes down as the frequency content of the background and the signal become more similar. These indications are supported by the predictions of the 2D and 3D Bayesian IOs which exhibit similar trends: the more similar the background and the signal frequency content, the smaller the difference between the 2D and the 3D task performance. Or, translated to the terms of the task difficulty, the greater the similarity between the frequency content of the background and the signal, the higher the difficulty of the task.

4.6.2 Experimental goal

The goal of our study is twofold: (1) to observe the trends in human detection performance in 2D versus 3D data sets, and (2) to investigate if (and how) the relationship between 2D and 3D human performance is influenced by the image properties (difficulty of the detection task).

4.6.3 Study design and methodology

An earlier study by [Burgess, 1999a], restricted to 2D image data, found the efficiency of humans relative to the 2D IO to be higher when correlation distances in the Gaussian-filtered noise background were less than Gaussian signal bandwidth (approx. 50%), and it dropped for correlation distances similar to the signal bandwidth (approx. 15%).

Here, given the fact that these are still early research steps in the area of the task-based evaluation of volumetric image quality, we decide to keep the task limited to signal detection (no search involved) and to use synthetic rather than clinical images in

order to be able to control the image data properties. Thus, we explore detection performance trends of human observers in a signal-known-exactly background-known-statistically task (SKE/BKE): the detection of a Gaussian signal (lesion) in a Gaussian lumpy background. We explore two parameters: the image viewing mode, single-slice (ss) versus multi-slice (ms) sequence-browsing image presentation, and the task difficulty determined by the frequency content of the background and the signal.

4.6.3.1 Image data

The images are synthesized as 3D correlated Gaussian noise with an added 3D Gaussian signal centered in the image volume. The correlated noise backgrounds are created by filtering 3D white-noise images with a 3D Gaussian shape filter. First, a background of size $196 \times 196 \times 63$ is extracted from a larger 256^3 field of view to avoid boundary effects. Then, in the z-direction, every three adjacent slices are averaged to simulate the slice thickness resulting in a $196 \times 196 \times 21$ data volume. The signal volumes are treated in the same manner. Finally, to one half of the backgrounds (signal-absent stacks of 21 slices), the signal of a given amplitude is added to create the signal-present images. The image data is created and the aforementioned imaging operations are performed in floating point 32-bit precision.

Table 4.4 summarizes parameter values of the synthesized images illustrated in Figure 4.21. We consider three different Gaussian noise kernels: $\sigma_{b1} = 11$, $\sigma_{b2} = 7$ and $\sigma_{b3} = 3$. The corresponding image categories are named B11, B07 and B03, respectively. For each background category, the peak signal intensity a_s is determined with the staircase method [Cornsweet, 1962, García-Pérez, 2001] using human observers, targeting the AUC of approximately 0.7 in ss mode. Additionally, a slightly higher signal amplitude is considered: $a_{s(i+1)} \approx 1.1 a_{si}$, $i \in \{1, 3, 5\}$. The exact same set of backgrounds from a given category is used with both a_{si} and $a_{s(i+1)}$ signal amplitude. In total, we explore six different image setups: B11-a_{s1}, B11-a_{s2}, B07-a_{s3}, B07-a_{s4}, B03-a_{s5}, and B03-a_{s6}. The signal spread is kept constant across all images ($\sigma_s = 5$).

As noted in Table 4.4, each background category is associated with a certain level of task difficulty: low (B11), medium (B07), or high (B03). Here, the measure of difficulty is defined as the relationship between the characteristics of the signal and those of the background: a low-difficulty task, where background lumps are noticeably larger than the signal; a medium-difficulty task, where background lumps are slightly larger than the signal; and a high-difficulty task, where background lumps are slightly smaller than the signal.

In ss mode, only the central slice of the volume is presented to the observer, while in ms mode all 21 slices are available.

Table 4.4: Image data parameters

Test setup	Bkgr σ_b	Signal σ_s	Signal a_s	Level of difficulty
B11-a_{s1}	11	5	60	Low
B11-a_{s2}			65	
B07-a_{s3}	7	5	120	Medium
B07-a_{s4}			130	
B03-a_{s5}	3	5	135	High
B03-a_{s6}			145	

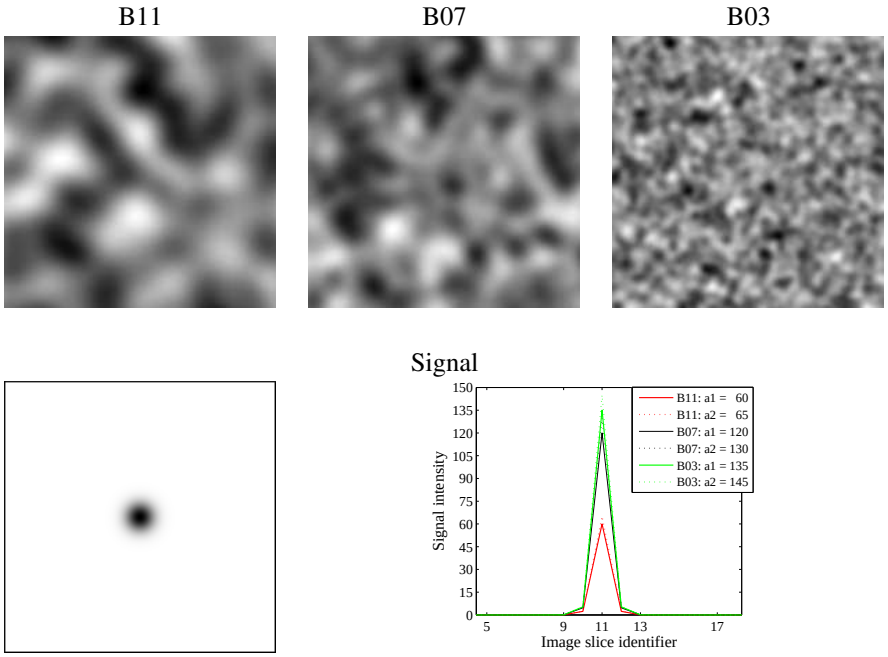


Figure 4.21: Experimental image data: (from left to right) one example image slice from each B11, B07 and B03 image background category; central slice of the signal image (for better visibility, the image intensities shown in the figure are inverted); and the plot of intensity profiles across central signal image slices.

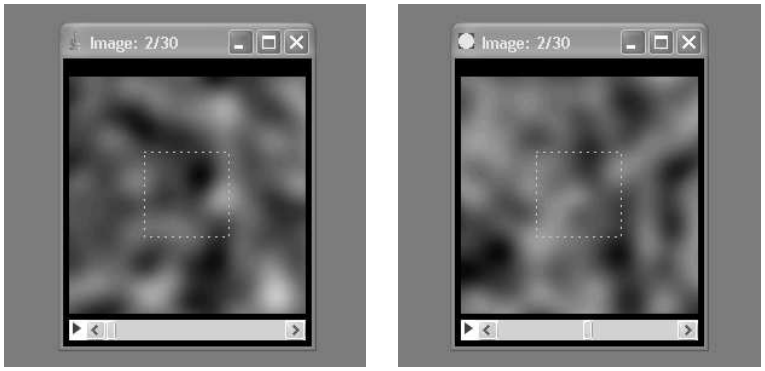


Figure 4.22: Graphical user interface used in the human observer study: (Left) multi-slice image presentation at the start of the sequence with indicated ROI in xy-plane (dotted rectangle), and (Right) multi-slice image presentation in the middle of the sequence, where the signal is expected as indicated by the white dot in the top left corner of the window.

4.6.3.2 Observer performance experiments

Twenty-two human observers took part in the MRMC ROC experiments. Two of them were very experienced with the tasks, two were moderately experienced (participated also in the pilot studies), and the others were newly trained. Two additional observers participated in the pilot readings only, including one expert neuroradiologist. Most participants were researchers in digital image processing, some had experience with medical image processing.

Human observers performed free inspection of a single stimulus (ss or ms image) and scored it using a 6 point confidence scale: definitely abnormal, probably abnormal, maybe abnormal, maybe normal, probably normal, definitely normal. As illustrated in Figure 4.22, the readers were aware of the approximate location of the target within the slice (ss and ms) and within the sequence (ms). With ms sequences, they were allowed to scroll through at arbitrary speed and direction. No time limitation was imposed.

The study consisted of four reading sessions, usually conducted on different days. The first session included training trials only, aimed exclusively at training the participants for the given task and not considered in the detection performance analysis. The testing trials were conducted in the subsequent three sessions. These were aimed at gathering test data – rating data of human observers which will later be used in the human performance analysis. Each session took approximately one hour to complete.

In training trials, the observer was given feedback after each trial. Also, when the classification decision was incorrect, the observer was allowed to repeat the trial in order to "learn from their mistake". Each observer did training trials with each background category and one of the two corresponding signal intensities, (see Table 4.4),

both in ss and ms mode. Approximately half of the 22 readers read images at signal level a_{si} and the other at level a_{si+1} (exceptionally, one reader read images at both their signal levels and one reader read only B07 and B11 images). The first reading session, the training session, involved 50 images of each of the three background types (25 normal and 25 abnormal cases) and each read in both ss and ms mode. Thus, the total number of images read during the training session was $50 \times 3 \times 2 = 300$.

Testing trials were split in three reading sessions, each dedicated to one background category (B11, B07 or B03) and involving both ss and ms readings. These trials were conducted for the same image parameters as those used in the training trials. Importantly, there was no overlap between the image sets used in training and those in testing trials. The test images were presented in a random fashion, grouped in ss and ms subsets. The number of testing trials per session was as follows: for B11, 64 ms + 64 ss; for B07, 84 ms + 84 ss, for B03, 94 ms + 94 ss. Each session contained an equal number of normal (signal-absent) and abnormal (signal-present) cases. The number of images per background category were chosen to allow the statistical significance in comparing mean AUCs for ms versus ss, based on the sample size estimates[20] from the pilot study analysis [Hillis and Berbaum, 2005, Hillis et al., 2005, Hillis, 2007]. The summary of MRMC human study parameters is included in Table 4.5.

In addition, to assess intra-reader variability (not reported here), in each experiment setup 6 images (3 signal-absent and 3 signal-present) were shown two times. Only the first human ratings of the repeated trials are considered in the performance analysis. Finally, to reduce variability in human-performance, 30 (B11) or 40 (B07, B03) trials with feedback preceded both ss and ms testing trials. Thus, a total number of trials per session was: 200 (B11), 260 (B07) and 280 (B03).

4.6.3.3 Image display

All images read by human observers were displayed on a Barco Coronis 5MP 10-bit grayscale digital mammography display calibrated to DICOM GSDF and with the luminance-response curve as shown in Figure 4.23. The native resolution of the display is 2048×2560 while the image area is 196×196 pixels or about 3 cm $\times$ 3 cm. The images were displayed in the center of the display and viewed on-axis. The software for displaying the images (loaded in floating point 32-bit precision) and collecting the observer responses was developed in form of a plugin for ImageJ program.[21] For an illustration of the graphical user interface see Figure 4.22. No image processing (*e.g.* zoom in/out, window/level adjustment) was allowed.

All human readings were conducted in a psychophysical test room [Marchessoux and Kimpe, 2007] shown in Figure 4.24 in order to ensure a controlled viewing envi-

[20] For more details, the reader is referred to the "Sample Size Estimation Overview" by the Medical Image Perception Laboratory of the University of Iowa, `http://perception.radiology.uiowa.edu/SampleSize/tabid/182/Default.aspx`.

[21] `http://imagej.nih.gov/ij/`

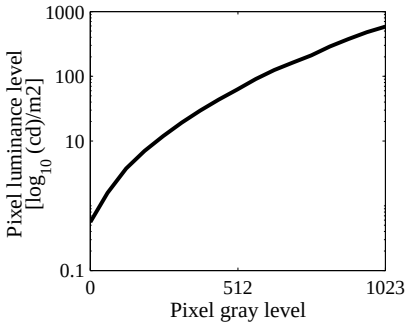


Figure 4.23: Luminance-response curve of the display used in the human observer study.

Figure 4.24: Psychophysical test room used in the human observer study.

ronment: fixed uniform and low intensity of ambient light and low surface reflectance (approximately 20%) which prevents glare effects. The readers were seated 50 cm from the display and they were allowed to lean back and forth, while the chair position was kept fixed. Before each reading session, the monitor was warmed up for at least one hour so that the luminance and the temperature of the white point become stable.

4.6.3.4 Performance measures

The rating data of human observers are analyzed using the MRMC ROC analysis of Dorfman, Berbaum and Metz (DBM) [Dorfman et al., 1992]. We use the DBM MRMC software, version 2.32 Build 3 [Dorfman et al., 1992, Hillis and Berbaum, 2005, Hillis et al., 2005, Hillis, 2007, Hillis et al., 2008], with readers and cases both treated as random effects in the ANOVA using AUC as the figure of merit. ROC curves are estimated using the non-parametric Trapezoidal-Wilcoxon estimation method and the error bars are 95% confidence intervals (CIs) on the AUC.

For the purpose of computing the relative efficiency of ss to ms mode performance, the AUC values are converted to the task SNR values using Eq/. (3.16). Finally, the relative human efficiency in ss versus ms mode is computed as:

$$\eta_{\mathrm{ss,ms}} = \frac{{\mathrm{SNR}_{\mathrm{ss}}}^2}{{\mathrm{SNR}_{\mathrm{ms}}}^2}, \tag{4.11}$$

where $\mathrm{SNR}_{\mathrm{ss}}$ and $\mathrm{SNR}_{\mathrm{ms}}$ stand for the SNR performance in ss and in ms image viewing mode, respectively.

4.6.4 Results

First, we look into the distribution of classification decision outcomes across different image presentation modes (ss and ms) and different test image setups (B11-a_{s1}, B11-a_{s2}, B07-a_{s3}, B07-a_{s4}, B03-a_{s5}, B03-a_{s6}). The cutoffs for correct and incorrect classification were based on the middle of the 6-point scoring scale. Four decision outcomes are possible: true positive (TP) for a correctly marked abnormal case and true negative (TN) for a correctly marked normal case; false negative (FN) for an incorrectly marked abnormal case and false positive (FP) for an incorrectly marked normal case.

Table 4.5 shows percentages of the total number of correct (TP+TN) and incorrect (FP+FN) classification decisions over all readers in a given test setup. These values indicate that the availability of additional slices in ms relative to the ss mode greatly reduces the number of incorrect classifications, more so for B07 and B11 image setups (lower-difficulty tasks). In particular, in the case of B03, the percentage of misclassified cases drops from about 30% in ss to about 20% in ms mode, while for B07 and B11 it falls from about 35% or 30% in ss down to approximately 5% or less in ms

Table 4.5: Human observer performance in multi-slice (ms) versus in single-slice (ss) mode

Test setup	MRMC $N_{\rm rd}$	MRMC $N_{\rm ts}$ Abnl		Decision outcomes TP+TN [%] Nl		Decision outcomes FP+FN [%] ms		Difference ms to ss ΔAUC ms	95% CI	$p-$value
B11-a_{s1}	11	32	32	68.32	96.45	31.68	3.55	0.27	(0.18, 0.36)	0
B11-a_{s2}	12	32	32	70.70	95.05	29.30	4.95	0.24	(0.15, 0.32)	0
B07-a_{s3}	11	42	42	66.56	93.51	33.44	6.49	0.26	(0.18, 0.33)	0
B07-a_{s4}	12	42	42	63.89	95.83	36.11	4.17	0.30	(0.22, 0.38)	0
B03-a_{s5}	12	47	47	66.40	77.13	33.60	22.87	0.14	(0.06, 0.22)	0.001
B03-a_{s6}	10	47	47	67.45	79.47	32.55	20.53	0.14	(0.07, 0.22)	0.001

presentation mode. These percentage values are only meant to give an indication of the human performance as we actually use the fitted AUC as the figure of merit.

Translated to the AUC domain, the contribution of additional slices in the ms over ss mode can be expressed in terms of the difference between AUCs in ms and ss treatments, $\Delta\mathrm{AUC} = \mathrm{AUC}_{\rm ms} - \mathrm{AUC}_{\rm ss}$. These values are presented next in Table 4.5 together with the corresponding 95% CIs, all obtained with the DBM MRMC analysis. In line with the trends suggested by the differences in ms versus ss percentages of incorrect classifications, the values of $\Delta\mathrm{AUC}$ are smallest for B03 image setups (0.14), and they become notably larger for B07 and B11 setups (≥ 0.24). Remember that the average AUC values in ss are by design approximately 0.7 for all test image setups. All reported differences are statistically significant with $p < 0.001$.

The AUC values of individual readers are shown in Figure 4.25. The three plots correspond to the three different background types considered in the study. In each plot, the horizontal axis represents values of per reader AUC in ms viewing mode while the corresponding AUC values of the same reader in ss mode are displayed on the vertical axis. By visually inspecting the three plots, we notice that for B11 and B07 backgrounds, the spread of per reader AUC values in ms mode is larger for B03 (approximately, AUC$\in (0.6, 0.9)$) than for B07 and B11 (approximately, AUC$\in (0.9, 1.0)$).

In Figure 4.26 we present the average AUC values of all readers per test image setup together with 95% CIs as error bars. In addition, Figure 4.26 shows the SNRs equivalents of human AUCs computed using Eq. (3.16). These SNR values are used in Eq. (4.11) to estimate the relative human efficiency in ss over ms image presentation mode: $\eta_{ss,ms}$, depicted in Figure 4.27. Again, given the smallest difference between ms and ss performance of humans in B03 setups, the ss to ms efficiency is largest in these data setups ($\approx 25\%$) and it drops in B07 and B11 setups ($\approx 10\%$ or less). In other words, the relative efficiency of ss to ms image presentation decreases with the decreasing level of task difficulty.

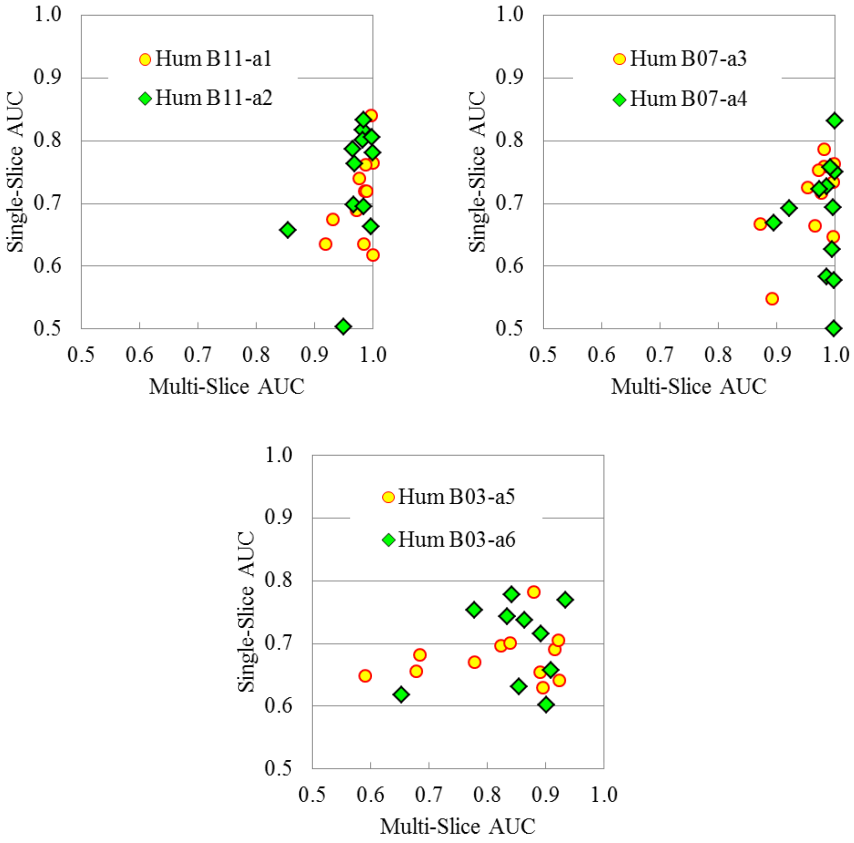


Figure 4.25: AUCs of individual human observers for ss and ms image presentation. The data is shown for all six experiment setups: B11-a_{s1} and B11-a_{s2} (left), B07-a_{s3} and B07-a_{s4} (middle), B03-a_{s5} and B03-a_{s6} (right).

In depth analysis of the human behavior parameters from this study are reported and discussed in [Kumcu et al., 2012b].

4.6.5 Discussion

As previously discussed by [Rolland and Barrett, 1992] and by [Burgess, 1999a] for 2D tasks and in Chapter 3 for 2D versus 3D tasks, the signal detection performance is influenced by the parameters of the image objects (background and signal).

First, we refer to the values of signal amplitude a_s chosen for each background setup B11, B07 and B03 based on the criterion of average human AUC of approximately 0.7 in ss viewing mode. As shown in Table 4.4, the signal amplitude required to reach AUC of 0.7 was highest for the backgrounds with lumps slightly smaller than the signal ($a_{s6} = 145$, $a_{s5} = 135$ for $\sigma_{b3} = 3$, $\sigma_s = 5$), it was slightly lower for the backgrounds with lumps slightly larger than the signal ($a_{s4} = 130$, $a_{s3} = 120$ for $\sigma_{b2} = 7$, $\sigma_s = 5$) and it was clearly the lowest for the backgrounds with the largest lumps ($a_{s2} = 65$, $a_{s1} = 60$ for $\sigma_{b1} = 11$, $\sigma_s = 5$). This result is consistent with the report by [Burgess, 1999a] who found the amplitude threshold (the value of amplitude required for a certain level of performance) to be the highest when the signal size and the correlation distance in the filtered noise background were approximately equal.

Burgess also pointed to the obvious consequence of this observation that detection is most difficult when the signal and the filtered noise have approximately the same spectral bandwidths (what we refer to as high task difficulty). This is in line with the vision literature on masking [Stromeyer and Julesz, 1972], which suggests that the degree of masking (difficulty in detecting the signal) increases with the increasing similarity in the frequency content of the signal (Gaussian blob) and that of the mask (background lumps).

Next, we refer to the study by [Rolland and Barrett, 1992] who found that, in 2D image domain, an increase in the mean number of background lumps and their strength are related to a decrease in the performance of humans. In order to characterize the background images used in our study, we will instead consider a few simple statistical properties which are commonly used to quantify image texture. Another way to relate the obtained results to properties of the backgrounds is to use image texture descriptors. Note that all the following measures can be computed even when the underlying statistical model of the data is not known, as is the case with real clinical images.

Table 4.6 presents the values of the measures we are interested in: mean and standard deviation of the background image pixel values, respectively, H-Mean and H-Std; and two standard statistical measures of image texture: mean local intensity range, and mean local standard deviation of the image, respectively, L-Range and L-Std. Here, the local intensity range of each pixel is determined by the difference between maximum and minimum value in its neighborhood (3-by-3), and its local standard deviation refers to the standard deviation in pixel intensity within the same neighborhood. The measures in Table 4.6 are computed as average values across all ss backgrounds of the

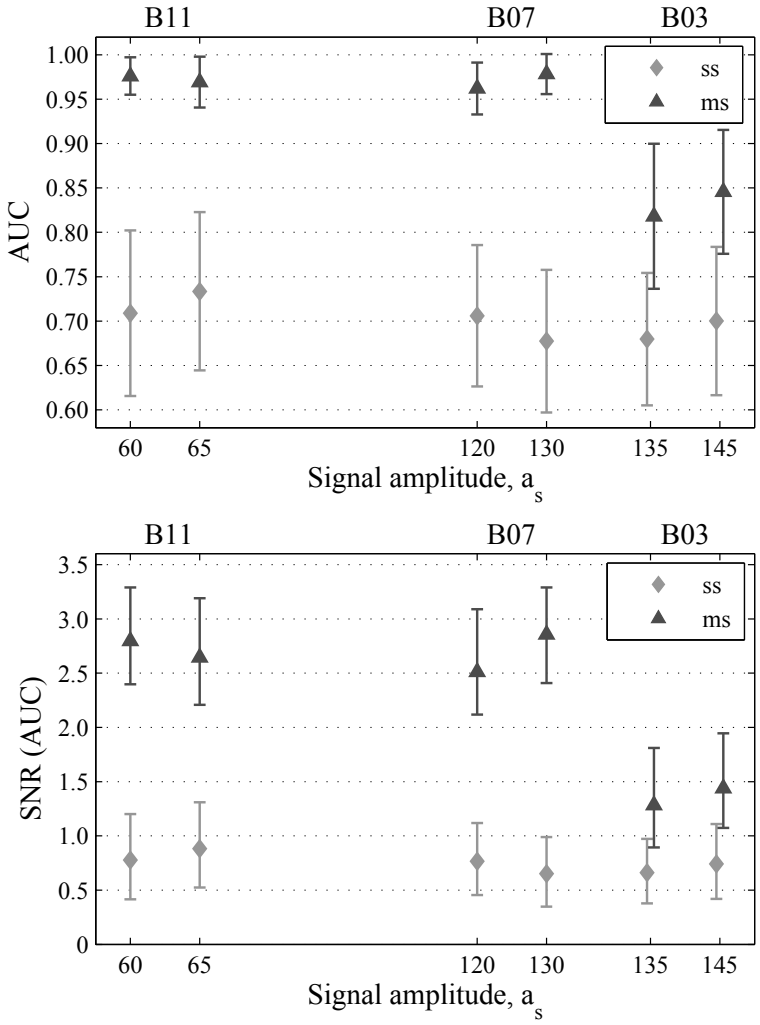


Figure 4.26: Average performance of the human observers in ss and ms image viewing mode: (left) AUC values obtained with the DBM MRMC analysis, and (b) SNR values computed using Eq. (3.16). The error bars indicate the 95% CIs for the AUC and their corresponding values in SNR domain.

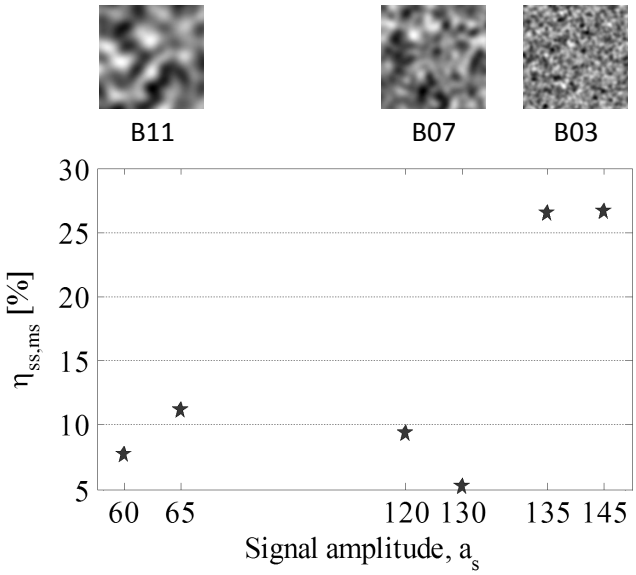


Figure 4.27: Relative efficiency of humans in ss versus in ms viewing mode: $\eta_{ss,ms}$.

same statistical distribution (B11, B07, B03).

The values of H-Mean and H-Std suggest that, while there are still some small differences among the three test setups, they are rather similar for the overall range of intensity values. Accordingly, the overall intensity (*i.e.* the corresponding luminance range) probably have no major impact on the performance levels for the corresponding detection tasks. On the other hand, the texture related measure L-Range indicates that the maximum to minimum intensity range in a local neighborhood is moderately different between B11 and B07, 31 and 44, but it is notably larger for B03 backgrounds, 85. A similar trend holds for the other texture measure L-Std, it is smallest for B11, slightly larger for B07, and largest for B03. Both of these suggest that the local texture of B11 is most smooth while B03 is most bumpy, which could contribute to higher signal detectability in B11 compared to B07 or B03.

These observations are in line with our assumption that the level of human detection performance could be tied not only to the degree of background-to-signal similarity in frequency spectra but potentially also to the strength of background texture. It will be of interest for future research to experimentally explore these assumptions. For example, we could consider two image parameter setups in addition to the present one, where the signal would be larger than in the present study and the size of the background lumps would be chosen such that the images have the same three ratios of the signal-to-background lump sizes as in our present study (see Table 4.4 for details). These experiments should reveal additional information about the actual relationship between the image data properties and the level of human detection performance.

Table 4.6: Background image statistics

Bkgr	Histogram		Texture (local)	
type	H-Mean	H-Std	L-Range	L-Std
B11	524	158	31	11
B07	513	139	44	15
B03	508	118	85	29

Finally, we offer some remarks regarding the effect of image properties on human performance in 3D detection tasks. As remarked by [Chen et al., 2002], ms images provide additional information which allows for a better distinction between true signals and noise or background structures. When comparing the 2D to the 3D detection performance of the models for Gaussian-filtered noise background with Gaussian signal as a target, our results in Chapter 3 indicated that the difference was smaller when the correlation size in the background and the size of the signal were more similar (when the frequency content of the signal and the background were more similar). The results for human observers obtained in our present study (see the values of ΔAUC from Table 4.5) suggest that the difference in AUC between ms and ss mode is larger for B07 and B11 compared to B03. Since the AUC performance in ss mode is approximately the same for all three types of the backgrounds, this translates to the fact that the benefit of using multiple slices is larger for B07 and B11 than it is for B03. Given the aforementioned trends captured by 3D model observers in relation to the image frequency content (background versus signal) and specific image parameters from our study summarized in Table 4.4, we would expect the increase in ms performance to be smallest for B03, larger for B07 and largest for B11. The actual trends measured for humans suggest that the ms performance of humans is affected not only by the level of similarity in the frequency content of the signal and the background but also by certain characteristics of the background data itself – in our case, the correlation length of the background. Our results suggest that for B03 backgrounds (smaller correlation size in the background) the additional slices fail to allow a much better distinction between the signal and the background structures in the z-direction due to smaller correlation size in the background. On the other hand, due to larger correlation size in B07 and B11 backgrounds, additional slices clearly improve distinction between the signal and the background structures in the z-direction which improves the signal detection performance in the ms image viewing mode.

4.7 Conclusion

In this chapter, we presented a series of observer studies directly or indirectly aimed at evaluating the utility of medical displays. The first of the four studies with model observers looked into the effects of different parameters of both the model itself and

the MRMC experiments. Similar to our results in Chapter 3, the results of this study suggest that the parameter values may significantly affect the results of the model observer studies, and thus it is of utmost importance that they are properly chosen and that the results are interpreted with caution and awareness of the associated limitations.

The subsequent three model observer studies explored the effect of the slow LCD response time on the detection performance in sequence-browsing mode. We have studied the msCHO performance for multi-slice images, either real clinical or computer-generated ones. The effects of image displaying at different frame rates have been simulated using two state-of-the-art models for the LCD temporal response. Overall, our results confirmed previous findings that the slow temporal response of medical LCDs degrades the detection performance of the observers – the higher the frame rate, the larger the degradations. Undoubtedly, this is a very important recommendation for the clinical practice: the rate of browsing through image volumes must be appropriately chosen (not too high) in order to avoid negative effects of the slow LCD temporal response, *i.e.*, in order not to introduce degradation in diagnostic accuracy. Importantly, our msCHO results suggested that the earlier estimates of the extent of these degradations by the ssCHO model could be overly pessimistic, *i.e.*, although evidently existing, the decrease in signal detectability caused by the slow LCD response time may be not as large and abrupt as predicted by the ssCHO.

One of the three msCHO studies of the LCD temporal response examined the benefit of a novel algorithm for compensation of the slow temporal response of medical LCDs. The results suggested improved detectability with the compensated LCD compared to a conventional one; the novel compensation algorithm was able to recover most of the degradation in signal detection performance caused by the slow LCD response time. Importantly, the results of this study were used as a preclinical validation of an actual display system. Moreover, the same msCHO experiments were able to correctly guide the parameters of the followup clinical study with human observers.

Next, for the purpose of more accurately assessing the effects of the slow LCD displays, we proposed an extension to the msCHO design, the upsampled msCHO model. Unlike the msCHO which considers only the end-of-frame on-LCD luminance values, the upsampled msCHO also addresses the within-frameon-LCD image information. Our results showed that integrating the within-frame information into the model observer allows it to be better aware of the LCD temporal luminance variations. Depending on the details of the luminance profile, neglecting the within-frame luminance information may lead to under- or overestimation of signal detectability.

Lastly, as suggested by several previous studies as well as by our results in Chapter 3, the data collected in our human observer study indicated that the level of human detection performance is influenced by image properties. In particular, when comparing single-slice to multi-slice, we found that the difference in performance increased as the task difficulty decreased. Thus, the benefit from additional image data in multi-slice mode is larger for lower-difficulty tasks. These results, together with further research on the mechanisms underlying the observed trends in human observer per-

formance (including, but not limited to, contrast sensitivity function (CSF) [Barten, 1999], temporal CSF [Barten, 1999], masking [Stromeyer and Julesz, 1972], internal noise [Abbey and Barrett, 2001,Zhang et al., 2007,Brankov, 2011]), shall aim to guide the design modifications to the msCHO model observer such that it can better predict detection performance of the human observers.

The work reported in this chapter already resulted in five peer reviewed international conference papers as first author [Platiša et al., 2009d, Platiša et al., 2010c, Platiša et al., 2011g,Platiša et al., 2011h,Platiša et al., 2012c] and another one as co-author [Kumcu et al., 2012b]. The publications also include four other abstracts and scientific conference presentations (three of which as first author) [Platiša, 2008, Platiša et al., 2010b, Platiša et al., 2011f, Kumcu et al., 2011c]. A journal article discussing the human observer study of single-slice versus multi-slice image viewing is in preparation [Platiša et al., 2014b].

5

Blur identification

This chapter researches the methods for the identification of image blur, the most common image distortion next to image noise. We start with introducing the models of image blur and briefly reviewing the basic principles of multiscale image analysis. Subsequently, we introduce a novel measure of image blurriness which relies on the ability of the wavelet transform to characterize edges in the image. The proposed measure can be used in both the full-reference (FR) and the more challenging and more realistic no-reference (NR) image quality assessment (IQA), *i.e.*, both with- and without the reference image (the golden truth). Furthermore, we formulate a novel edge descriptor and explain how it can be applied to the problem of edge-based image matching. Finally, we perform an extensive comparative performance analysis involving a number of the state-of-the-art techniques for the NR image blur assessment.

5.1 Introduction

Digital image blur is one of the most common causes of image quality (IQ) distortion. In some of the principal standard dictionaries online *blur* is defined as "something vaguely or indistinctly perceived"[1] or "something that you cannot see clearly" [2]. In technical terms, we often refer to the concept of *edge* structures in the image and describe blur as loss of *sharpness* of the edges, where a sharp edge assumes a step discontinuity in image pixel values. Hence, the terms blurriness and sharpness are used to refer to the same property of images, only in reverse proportion, *i.e.*, more blurriness implies less sharpness and blur-free suggests perfectly sharp. Alternatively, the term *unsharp* is also used in the literature to mean blurry.

Commonly, it is already at the stage of digital image acquisition that digital image blur occurs, often together with image noise. Overall, we can distinguish two categories of causes of blurriness: the *optical* and the *control-related* factors, both having an impact on our perceived IQ. Discussing properties of the boundaries of physical

[1] http://www.merriam-webster.com/dictionary/blur

[2] http://dictionary.cambridge.org/dictionary/british/blur_1?q=blur

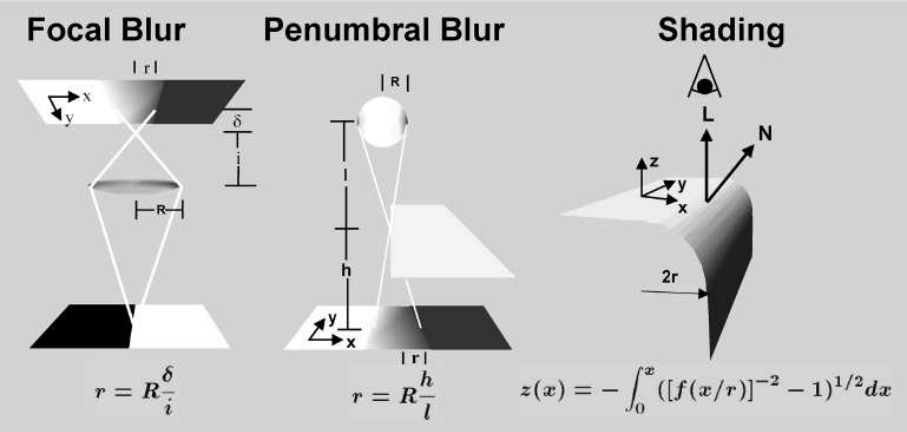


Figure 5.1: Edges in the world generically project to the image as spatially blurred (from left to right): focal blur due to finite depth-of-field; penumbral blur at the edge of a shadow; shading blur at a smoothed object edge. Figure taken from [Elder and Zucker, 1998].

structures in the world, [Elder and Zucker, 1998] point to the fact that – due to the optics of the camera – these generally project to the image as spatially blurred (rather than perfectly sharp). Figure 5.1 illustrates the scenarios which cause such unsharp appearance of object boundaries in the images: the finite depth-of-field, an imperfect light source resulting in occurrence of shadows, or simply a rounded (rather than straight) edge of the actual imaged object. These are what we refer to as *optical* causes of blur in the images.

Another set of common causes of blurriness, here referred to as *control-related* factors [Yoshida, 2005], has its origin in the way the camera is handled, either by a human (*e.g.* the photographed object is not well focused, exposure-time is long when the object is moving, camera shake) or by an automated camera control system (*e.g.* auto-focusing error).

Systematic reviews of different natures of blur regularly encountered in practical situations of interest and commonly considered in the research efforts can be found in [Wang and Bovik, 2006, Lagendijk and Biemond, 2009]. In this work, we investigate three most common blurs: *linear motion*, *atmospheric turbulence* and uniform *out-of-focus* (hereafter defocus) blur. For details about these three blurs and their mathematical models, we refer to Section 5.2.

The other most common type of image distortion, next to image blur, is image noise. Similar as blur, the main noise is induced already in the acquisition process (*e.g.* sensor noise, dark noise, *etc.*). For the purpose of our investigations we consider the widely used additive Gaussian white noise model. For more in depth details about origins of noise and advanced noise modelling the reader is referred to the literature

[Holst and Lomheim, 2011, Lagendijk and Biemond, 2009, Fiete, 2010, Nakamura, 2006].

Estimating the amount of blurriness, or the level of blur (BL) no matter what type, is of crucial importance for various imaging and image processing tasks. Moreover, for many applications it is essential that this estimation is *not* influenced by other kinds of distortions, such as noise. For example, image based passive auto-focusing algorithms use sharpness measures as the criterion function to find the focus position [Yao et al., 2006]; image restoration algorithms rely on estimates of the distortion parameters to perform image deblurring and denoising; some state-of-the art objective IQ schemes start by quantifying individual distortions and perform regression analysis (or factor analysis) to deduce the overall IQ scores. In the area of IQ assessment (IQA), especially perceptual IQA, it is necessary to relate visual impressions to accurate quantitative measurements of blurriness – this is important not only for theoretical psychovisual studies but also for practical video distribution systems where it is usually needed to compromise between different image distortions such that the visual experience of an end user is maximized [Papp et al., 2009].

Commonly, we differentiate between three scenarios in which the image distortion can be assessed: a *full reference* (FR) scenario in which the non-distorted, hereafter the reference (REF), image is also available; a *reduced reference* (RR) scenario where only restricted information about the REF image is allowed; and a *no-reference* (NR), also called *blind assessment* scenario in which the distortion-free image is completely unavailable. Being the most frequent scenario in real-life applications, the focus of our investigation is NR scenario. In this chapter, we propose a new technique for identification of image blurriness which unlike a majority of the existing state-of-the-art NR blur measures succeeds in making independent estimates of BL even in the presence of very high noise distortions.

The key contributions of the work reported in this chapter are the following: (1) we introduce a novel method for NR blur identification, (2) we propose a novel edge descriptor, and (3) an algorithm for image matching (image similarity measure) based on the edge content of the images. Our results demonstrate that the proposed NR blur measure is able to identify different types of blur (Gaussian, motion, defocus), while most of the existing measures perform well only for some specific blur types. We test and compare performance of twelve state-of-the-art NR blur measures (inclusive the proposed) for three types of blur: Gaussian (GBlur), motion (MBlur) and defocus (DBlur); both in the absence of noise and in varying levels of additive Gaussian white noise (GWN).

Further in this chapter, in Section 5.2, we present the models of image blur and illustrate the effects of the three different blur types considered in our work. The main concepts around multiscale image analysis and the key benefits of wavelet image decomposition for the purpose of edge characterization are presented in Section 5.3, followed by a brief description of the wavelet-based measure named *average cone ratio* (ACR) which is the basis for our proposed algorithms, introduced next. First,

the novel NR blur measure called CogACR is introduced in Section 5.4. Next, in Section 5.5 and Section 5.6, respectively, we describe the novel ACR-based edge descriptor and how it can be used in an image dictionary matching scheme for NR blur identification. In Section 5.7, we review the basic concepts of a number of state-of-the-art NR blur measures which will be also used in the comparative analysis of methods for NR blur identification. Those and other experimental results are presented and discussed in Section 5.8. Finally, Section 5.9 concludes this chapter.

5.2 Digital image blur models

Scientific study of the real world processes behind image blurriness and development of numerical algorithms for objective characterization of the extent of blur require mathematical formulations of the problem: the image formation process.

We denote by $f(x, y)$ a two-dimensional (2D) image which is free from any (technical) distortion (blur and noise included); in the context of image restoration. This distortion-free image is commonly referred to as an "ideal" image, while in the IQA context, we often refer to it as the "reference" image, REF. In general, when the image $f(x, y)$ is corrupted by blur and contaminated by random noise, the degraded image $g(x, y)$ can be described as

$$g(x, y) = f(x, y) * h(x, y) + n(x, y). \tag{5.1}$$

Here, * denotes the 2D linear convolution, $h(x, y)$ is the convolution kernel, also known as point spread function (PSF) or simply the blurring function that acts on an "ideal" image, and $n(x, y)$ is the noise contribution at the corresponding spatial position (x, y). Throughout the chapter, the noise is assumed independent of $f(x, y)$ and defined as zero-mean *additive white Gaussian noise* with the standard deviation σ_n, $n \sim N(0, \sigma_n)$ (hereafter referred to as GWN, or simply noise).

Another common simplification in blur identification studies (ours included) is spatial invariance of blur, meaning that the image is blurred in exactly the same way at every spatial location [Lagendijk and Biemond, 2009]. Clearly, this is in contrast with the many real cases in which the blur is varying across image area, either in its origin or in extent, or both (*e.g.* different focus of foreground and background, motion blur occurring only for moving objects in the scene). One obvious way to evaluate spatially varying blurs is to perform *local* (*e.g.* per image block) rather than *global* blur identification. Thereby, we allow for spatially varying blur characteristics to be adequately captured by the measure.

Figure 5.2 depicts example blur kernels used in our experiments.[3] Additionally, the effects of applying these kernels on four basic types of edges encountered in nat-

[3] All illustrations and experimental results reported in this chapter are based on computer-simulated image distortions (blurring and adding noise) performed using Matlab. First, blurring filters were created using `fspecial()` with the `'type'` parameter set to `'gaussian'`, `'motion'` or `'disk'` (for DBlur), and applied on reference grayscale images using `imfilter()`. Finally, these blurred images were added Gaussian white noise generated with `imnoise()`. The simulation process corresponds to that

ural scene images as defined by [Tong et al., 2004] (more details are provided in Section 5.7) are illustrated in Figure 5.7. Note that, unless explicitly stated otherwise, the examples and illustrations in this chapter assume DBlur.

5.2.1 Gaussian blur, GBlur

In the literature, the most frequently explored model of blur is Gaussian function which has been shown to describe reasonably well the blur introduced by atmospheric turbulence, especially under long-term exposures [Lagendijk and Biemond, 2009]. Using σ to denote the amount of spread of the blur and C a constant, the PSF of GBlur is given by

$$h(x, y) = \eta \exp \frac{-(x^2 + y^2)}{2\sigma^2}. \tag{5.2}$$

Though it may be not the very best representation of the blur encountered in real natural scene imagery, this model is widely used in the research domain. On the one hand, Gaussian PSF allows comprehensive analytical exploration for the purpose of image restoration. On the other hand, several public image databases of Gaussian blurred images exist for which perceptual IQ scores have been gathered which allows new perceptual IQ methods to be comparatively assessed on a real human data. More details about such databases can be found in [Winkler, 2012].

5.2.2 Defocus blur, DBlur

Assuming circular camera aperture, the DBlur can be modelled by a circular averaging filter (pillbox) within the square matrix of $2r + 1$ pixels in size, where r is the radius of blur:

$$h(x, y) = \begin{cases} 1/(\pi r^2), & \sqrt{x^2 + y^2} \leq r \\ 0, & \text{otherwise} \end{cases}. \tag{5.3}$$

5.2.3 Motion blur, MBlur

As elaborated by [Cai et al., 2012], the diversity of possible causes of MBlur (for example, translational or rotational or combined fast movement of an object in the scene, slow movement of an object under a long exposure time, camera shake) and its manifestations (for example, a smeared moving car versus a smeared whole image), makes MBlur modelling rather different and notably more complex then the previous two blurs, GBlur and DBlur.

Instead, the model is often confined to an important case of global translation. This occurs, for example, in the case of a camera shake while the scene being photographed

from the ASU Image and Video QUality Evaluation SofTware[4] (IVQUEST) [Murthy and Karam, 2010], also used in an extensive study of video quality measures by [Chikkerur et al., 2011]. The only difference from IVQUEST is that, rather than the fixed 3×3 default Matlab size, the size of GBlur kernel was set to be a square of the size $3\sigma_n$ (rounded off), the same as in [Jayaraman et al., 2012, Soleimani et al., 2013]; for MBlur and DBlur the default Matlab parameters were used.

(a) GBlur ($\sigma = 5$) (b) MBlur ($l = 5, \theta = \pi/4\text{rad}$)] (c) DBlur ($r = 5$)

Figure 5.2: Blur kernels (PSFs) of the three blur models used in our study: GBlur, MBlur and DBlur. The kernel matrices correspond to the mid amounts from the considered ranges of blurriness.

is static. Assuming constant velocity of the translation $v_{relative}$ under an angle of θ radians with the horizontal axis relative to the camera, the length of motion is $l = v_{\text{relative}} t_{\text{exposure}}$ and the MBlur kernel can be represented as

$$h(x,y) = \begin{cases} \frac{1}{l}, & \sqrt{x^2+y^2} \leq \frac{l}{2} \text{ and } \frac{x}{y} = -\tan\theta \\ 0, & \text{otherwise} \end{cases}. \tag{5.4}$$

5.3 Multiscale image analysis

Digital image analysis (by computers) as a counter part for visual image interpretation (by humans) is aimed at extracting certain information from the image data. In fact, the range of applications is rather staggering, and yet new use cases are continually emerging. Each use case is identified by a specific *task* for the analysis [Romeny, 1996, Lindeberg, 1996], ranging from rather simple ones such as deciphering the bar coded price tags in a supermarket, to more sophisticated tasks such as identifying a person from their face, or counting people to ensure that the building is below the safe level of occupancy, or delineating subtle cortical lesions in the images of human brain.

While we know that humans are very good at interpreting visual information, the essence of how they do that is still far from being understood. [Witkin and Tenenbaum, 1983] write: "We impose *organization* on data (noticing flow fields, regularity, repetition, etc.) even when we have no idea what it is we are organizing. ... the naive observer often sees essentially the same thing as an expert does. ... It is almost as if the visual system has some basis for guessing *what* is important without knowing *why*." Hence, digital image analysis can not directly rely on the (low-level) principles of the human visual system (HVS), and instead has to come up with its own ways for extracting the desired information from images.

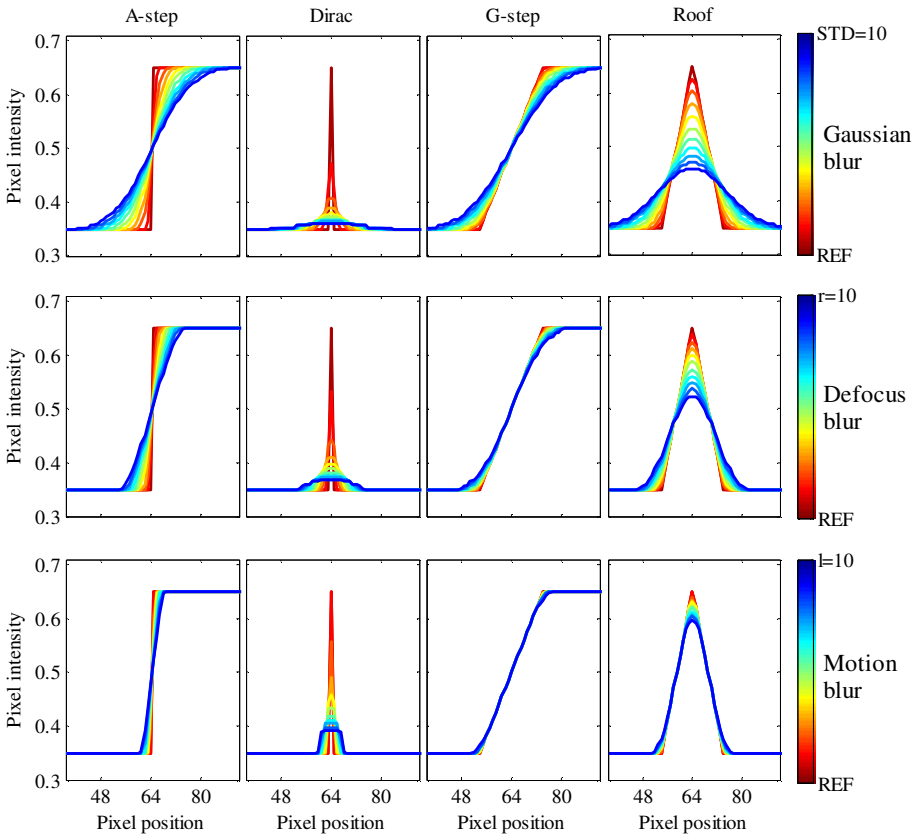


Figure 5.3: Influence of blur on four different types of edges [Tong et al., 2004] (from left to right): A-step structure, Dirac structure, G-step structure, and Roof structure. Each graph depicts pixel values of an undistorted edge (REF) as well as the values of that same edge when distorted by different types and extents of blur: (top) GBlur, $\sigma \in \{1, \ldots, 10\}$; (middle) MBlur, $l \in \{1, \ldots, 10\}$, $\theta = 45°$; (bottom) DBlur, $r \in \{1, \ldots, 10\}$.

Digital images are conventionally represented in *pixel domain* where each pixel value encodes a measurement of light reflected from a spatially corresponding surface area of the imaged physical scene. Assuming the simplest case of a single spectrum data (also referred to as monochromatic or grayscale), pixel value contains only intensity information encoded using certain pixel depth (typically 8 bits per pixel (bpp) in general purpose images, 10 bpp or more for specialized applications such as medical imaging or remote sensing).

For humans to view a digital image, the matrix of numerical pixel values is commonly visualized (transformed, typically in a non-linear fashion) into the matrix of "dots" colored in shades of gray (*e.g.* 512×512 gray shade dots) – either by printing or by showing the image on an electronic display device. Once the image is visualized in the *"gray shade domain"*, the HVS is usually very successful in interpreting (extracting information from) the image data.

However, for computer-based image analysis, especially for automated analysis (*e.g.* noise suppression, compression), the representation based purely on intensity values and spatial position of the pixels is not the best suited one. Commonly, we are interested in "organizing" the data and analyzing structure-related information in the images (lines, contours, shapes, etc.). Automated algorithms often do that by analyzing the local variations of the image intensity [Mallat, 1989] and these can easily get disturbed if the image is represented purely as a collection of dots. Consider for example the problem of edge detection in a noisy image. While humans will have no difficulty performing this task (even at very large amounts of noise), a computer algorithm which operates in the pixel domain (for example a well-known Sobel edge detector) may suffer from serious negative effects of noise; see Figure 5.25 for an illustration. Therefore, for the purpose of computer image analysis, we are interested in alternative image representations.

In the current state-of-the-art, the class of *multiscale* image transformations is being most actively researched and continually expanded [Daubechies, 1988, Mallat, 1989, Donoho, 1999, Candès et al., 2006, Guo and Labate, 2007]. Central to the paradigm of multiscale image representation is the observation that human perception of objects in the real world as well as in the images depends on the scale of observation [Marr and Nishihara, 1978, Witkin and Tenenbaum, 1983, Koenderink, 1984], or the size of the object [Rosenfeld and Thurston, 1971]. This is often illustrated with an example of a tree: viewed from a large enough distance, the tree may appear as small and simple as a dot; then gradually reducing the viewing distance would result in the same tree being perceived as an increasingly larger round shape until we would start seeing the branches and the leaves; eventually, arriving at a small enough distance, we may be even able to see the fine texture of the leaves [5]. For an illustration in terms of practical applications, consider for example the aforementioned tasks of people count-

[5] Thereby, the concept of scale can be related (in an abstract way) to the concept of viewing distance which is a very important and carefully regulated parameter of human image viewing, especially in highly specialized applications such as diagnostic medical image inspection.

ing versus face-based identification – the scale required for "face counting" is rather coarse compared to the scale required for measuring "facial microgeometry". This clearly illustrates the role and importance of scales in image analysis.

The key feature of multiscale (or multiresolution) representation of image data is exactly the ability to represent the input pixel domain data at multiple scales, where the finest scale captures the finest details in image structure (corresponding to the smallest viewing distance) and the coarser scales aggregate details into larger simplified structures (corresponding to the growing viewing distances). In our work, we exploit the multiscale property of wavelet image decomposition. In particular, the inter-scale dependencies of wavelet coefficients allow successful edge detection and blur estimation while at the same time keeping the influence of noise very low. The details are presented next.

5.3.1 Wavelet decomposition

Typically, for the purpose of multi-scale image (signal) analysis, images are often represented in the wavelet domain. Essentially, the wavelet transform is determined by a function which is commonly referred to as the mother wavelet and denoted by $\psi(t)$, where t is a real variable (often time or space) [Daubechies, 1992, Mallat, 1999]. A mother wavelet is a function that can be characterized as follows: it oscillates like a wave[6] in a small interval of t, it is well localized (approaches zero outside this small interval), it has finite energy $\psi(t) \in L^2(\mathcal{R})$ and zero mean value $\int_{-\infty}^{\infty} \psi(t)\,\mathrm{d}t = 0$. In general, we generate a family of wavelets by dilating (stretching, changing the scale of) and by translating (changing the position of) the mother wavelet. If we use a and b to denote, respectively, the scale and the position, the wavelet function can be described as

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \left(\frac{t-b}{a} \right),\ a, b \in \mathcal{R}; \tag{5.5}$$

where $\frac{1}{\sqrt{a}}$ is the normalization constant introduced to ensure constant energy of the wavelet. Formally, the continuous wavelet transform (CWT) of a signal of finite energy $f(t)$ is defined as

$$\mathcal{W}f(a,b) = \int_{-\infty}^{\infty} f(t)\, \overline{\psi_{a,b}(t)}\, \mathrm{d}t = \langle f, \psi_{a,b} \rangle, \tag{5.6}$$

where $\overline{x}$ denotes the complex conjugate of x and $\langle \cdot \rangle$ is the inner product. Therefore, applying the CWT to the signal actually means analyzing correlations between the signal and the wavelets: translated and dilated versions of the mother wavelet. The coarse features in the signal get uncovered through correlations with more stretched (larger scale) wavelets while the fine signal features get explicit from correlations with

[6] The name *wavelet* is formed from the root *wave* and a diminutive suffix *–let* to mean a small wave.

small scale wavelets; hence also the names "coarse" and "fine" for large and small wavelet scales, respectively.

Based on Eq. (5.6), two obvious properties of the CWT are high redundancy and *shift-invariance*. For practical reasons, the scale parameter is often chosen from the dyadic sequence $a = 2^j, j \in \mathcal{Z}$, thereby the name *dyadic* CWT for $\mathcal{W}f(2^j, b)$. In terms of applications, the dyadic CWT has been extensively used for characterization of singularities in signals. In particular, [Mallat and Hwang, 1992] proved that the CWT magnitude can detect all the singularities of $f(t)$ and proposed strategies to quantify those singularities in terms of Lipschitz regularity (the concept is described in Section 5.3.3).

Even more practical is the case where both the scale parameter a and the position parameter b take discrete rather than continuous values. Typically, the position is sampled ("decimated") proportionally to the scale $b = k2^j$, $k, j \in \mathcal{Z}$. Correspondingly, the related wavelet transform $\mathcal{W}f(2^j, k2^j)$ is referred to as the *decimated discrete* wavelet transform (DWT).

In engineering terms, the DWT can be seen as a two-channel filter bank comprised of a scaling (low-pass) filter $\mathbf{h}$ and a wavelet (high-pass, or bandpass) filter $\mathbf{g}$, each followed by downsampling by factor 2. Conveniently, coefficients of coarser scales are computed from coefficients of finer scales by recursively applying the following equations on the low-pass output $\mathbf{s}_j$:

$$s_{j+1,k} = \sum_{l \in \mathcal{Z}} h_{2k-l}\, s_{j,l} \tag{5.7}$$

$$w_{j+1,k} = \sum_{l \in \mathcal{Z}} g_{2k-l}\, s_{j,l} \tag{5.8}$$

One decomposition step of the two-dimensional decimated DWT is illustrated in Figure 5.4. Notice the notation used to denote the four different subbands HH, HL, LH, and LL, and the corresponding filter outputs: three detail images $\mathbf{w}^{\mathrm{HH}}$, $\mathbf{w}^{\mathrm{HL}}$, $\mathbf{w}^{\mathrm{LH}}$, and the lowpass image $\mathbf{s}$. Commonly, elements of the detail images are referred to as the *wavelet coefficients*.

The advantage of providing a non-redundant representation of the signal makes the decimated DWT especially attractive for real-time and memory-constrained applications such as image/video compression (*e.g.* JPEG 2000 and Motion JPEG 2000). Nevertheless, the lack of shift-invariance makes it less desirable for applications involving statistical modelling, such as image analysis [Mallat, 1996] and image denoising [Coifman and Donoho, 1995].

For our work, we use the *non-decimated* DWT $\mathcal{W}f(2^j, k)$ in which the signal (in our case, the image) is represented by the same number $k \in \mathcal{Z}$ of wavelet coefficients at each scale $2^j, j \in \mathcal{Z}$. The transform is implemented using the *á trous* algorithm [Holschneider et al., 1989]. Practically, this means stretching the filters $\mathbf{h}$ and $\mathbf{g}$ in Eqs. (5.7) and (5.8) by inserting $2^j - 1$ zeros between each two of their coefficients. If we denote the stretched filters by $\mathbf{h}^j$ and $\mathbf{g}^j$, respectively, then the *á trous*

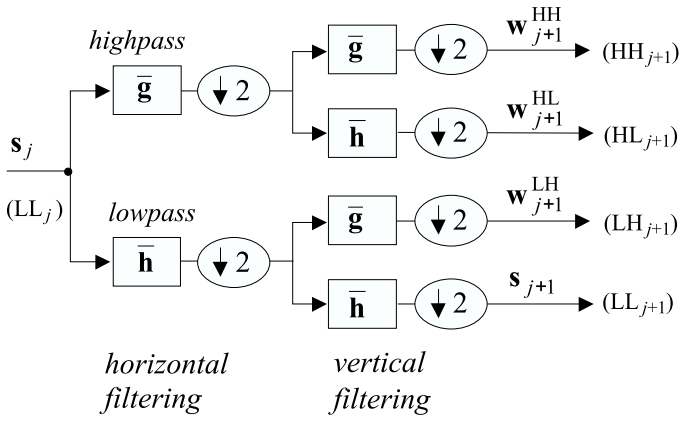


Figure 5.4: One decomposition step of decimated two-dimensional DWT. Figure taken from [Pižurica, 2002].

algorithm is

$$s_{j+1,k} = \sum_{l=-\infty}^{\infty} h^j_{k-l}\, s_{j,l} \tag{5.9}$$

$$w_{j+1,k} = \sum_{l=-\infty}^{\infty} g^j_{k-l}\, s_{j,l} \tag{5.10}$$

$$s_{j,k} = \frac{1}{2}\left(\sum_{l=-\infty}^{\infty} h^j_{k-l}\, s_{j+1,l} + \sum_{l=-\infty}^{\infty} g^j_{k-l}\, w_{j+1,l} \right) \tag{5.11}$$

Figure 5.5 illustrates the result of a 4-level non-decimated wavelet decomposition applied on a simple test image contaminated with different levels of DBlur and GWN. Note that the fine features coming from GWN are predominantly present at finer scales while at coarser scales these details gradually "disappear". Nevertheless, what is perhaps less obvious from this example, also some fine edges may get filtered out at larger wavelet scales, allowing only the more "rude" features to prevail (the silhouette of a head with a hat). Related properties of the wavelet transform will be further discussed in the following Section 5.3.2 and later in Section 5.3.3 as we explore the parameters for our proposed methods.

5.3.2 Robust edge detection for blur identification

We have established by now that image blurriness, or unsharpness, is determined by the properties of image edges. Accordingly, one often exploited strategy for blur identification is *characterizing edges*, which starts by edge detection.

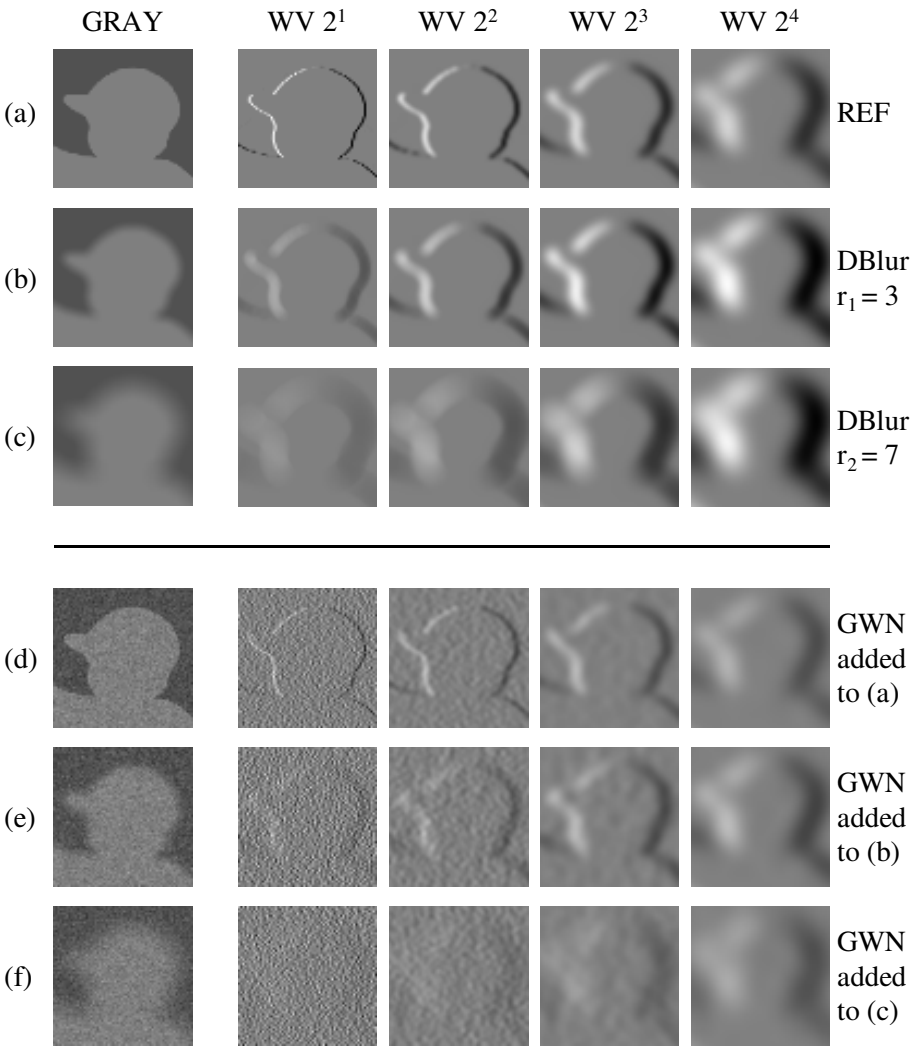


Figure 5.5: The first four scales of WT for: (a)-(c) different levels of DBlur (none, $r_1 = 3$, $r_2 = 7$) introduced to the REF image, and (d)-(f) images from (a)-(c) with added GWN of $\sigma_{n1} = 10$.

Commonly, the positions of detected edges are represented by a binary mask, hereafter referred to as an *edge map*. The map $\mathbf{e}$ is a set $\{e_1, \ldots, e_L\}$ of binary labels:

$$e_l = \begin{cases} 1 & \text{if } l \text{ is an edge pixel} \\ 0 & \text{otherwise} \end{cases} \tag{5.12}$$

where index $l = 1, \ldots, L$ represents image pixels in a raster-scan order (*e.g.* for a 256×256 pixels image, $L = 256^2$).

Clearly, a very important consideration in the process of edge detection is related to image content. Detecting exactly those edges in the image which are *relevant* for the analysis at hand is a challenging problem in itself, even if the image is not corrupted by noise or other distortions. For the purpose of blur estimation, we are often interested in identifying the dominant edges in the image and excluding those comparatively smooth or subtle (such as, for example, the edges within the bottom area of grass in the "Plane" image from Figure 5.7). Moreover, the edge map should omit the details from very high frequency textures in the image since these get distorted already at very low levels of blur and may be misleading for the judgment of overall blurriness of the image (for example, see the edges around points of the cactus in the "Cactus" image in Figure 5.20).

Conventional edge detection methods, such as Canny or Sobel edge detector, rely on image gradient magnitude: a point (image pixel) is classified as an edge pixel if its gradient is a local directional maximum and is above a certain threshold. The methods differ in the choice of smoothing filters and the particular way of computing the edge strength. However, the drawback of this approach is its sensitivity to common image artifacts such as blur or noise; refer to Figure 5.25 for illustration.

In contrast to the edge detection schemes operating in the pixel domain, we turn to wavelet decomposition of the image to determine the edge positions. As discussed for Figure 5.5, larger scales of wavelet transform are able to filter out noise reasonably well, even at high levels. Moreover, [Mallat and Hwang, 1992] provide a mathematical description of the difference between edge and noise singularities. Therefore, we expect an adequately designed wavelet-based edge detection scheme to exhibit a much higher immunity to noise compared to the existing intensity-based techniques. In the following, we describe the proposed wavelet-based algorithm for edge detection.

For illustration, let us first consider the case where an image is free from noise distortions. Commonly, the edges are detected by examining the detail wavelet images and identifying the positions of important (largest) wavelet coefficients, *e.g.*, by thresholding the coefficient magnitudes. Figure 5.6 illustrates several strategies of wavelet magnitude thresholding. Four test images are considered: the distortion-free images of "Houses" and "Peppers" (on the left) and their distorted variants with added GWN of $\sigma_n = 10$ (on the right). The corresponding edge maps shown in rows 2 and 3 are obtained by thresholding magnitudes of wavelet coefficients at *scale* 2^1 while, respectively, using a *fixed* threshold value for all images and using an *image-specific* threshold value (in this case, choosing 5% of the highest coefficient magni-

tudes). Keeping the threshold fixed ensures that all extracted edges are of the similar "strength" but, as can be seen from the edge maps in row 2 of Figure 5.6, this may result in large variability in the level of details depicted by the edge maps. In that sense, for the noise-free images at least, the strategy of choosing a fixed percent of the highest coefficients depicted in row 3 seems a preferred stategy. Apparently, in the presence of image noise, the 5% of the highest coefficients seems not the best choice – while the edge map for the "Houses" image seems reasonably good, for the "Peppers" image there is a lot of noise in the edge map, which is certainly an undesired effect.

As we saw earlier, larger wavelet scales are less affected by image noise. Therefore, in order to improve edge detection for noisy images, we apply the aforementioned two detection strategies on wavelet coefficients at *scale* 2^3. Remember from Section 5.3.1 that also the fine edges gradually disappear at the larger scales, and for this reason we avoid the very large scales. These results are shown in rows 4 and 5 of Figure 5.6, where the thresholding parameters are the same as for the rows 2 and 3, respectively. As expected, working at a larger scale results in improved edge maps for both thresholding strategies: the edge maps now better represent the outlines of the objects in images and are much less affected by noise. The "Houses" edge map is now clean from the many tiny edges and, as desired, it mostly contains strong and long edges. The only criticism of these results relates to the nearly undetected edges of the darkest house in the image (although indeed those edges are of the smallest strength, *i.e.*, the corresponding intensity transitions are the smallest among the different houses). Even more improvement is observed for the "Peppers" edge maps obtained at scale 2^3. In the case of fixed threshold value (compare row 4 versus row 2), we notice that more of the visually significant edges have been detected at scale 2^3. This aspect is further improved in the case of adaptive threshold (compare row 5 versus row 3) while at the same time the effects of noise are greatly suppressed – though still not completely eliminated (*e.g.* note in the edge map the tiny structures in the area of far left pepper and those around the stem of the central bell pepper).

As shown in the literature, edges and noise can be better distinguished in the products of (adjacent) scale coefficients than in the coefficients of a single scale [Sadler and Swami, 1999, Zhang and Bao, 2002, Bao et al., 2005]. In particular, for the case of GWN, [Mallat and Hwang, 1992] prove that the average number of local maxima decreases by factor 2 from one scale 2^j to the next 2^{j+1}. In contrast, the modulus maxima of edges propagates to larger scales. Accordingly, the scale multiplication magnifies edge coefficients and suppresses noise.

With this in mind, we examine also the performance of edge detection by thresholding an intermediate detail image acquired by multiplication of wavelet detail images at different adjacent scales. We use $\mathrm{P}_{n\to k}$ to denote an inter-scale product of wavelet coefficients at dyadic scales $[2^n, 2^k]$ where $k \geq n+1$. Prior to scale multiplication, possible shifts introduced in the detail images by wavelet transform shall be compensated [Mallat and Zhong, 1992]. Based on the previously discussed experiments with single scales, se well as several reported methods using multiscale prod-

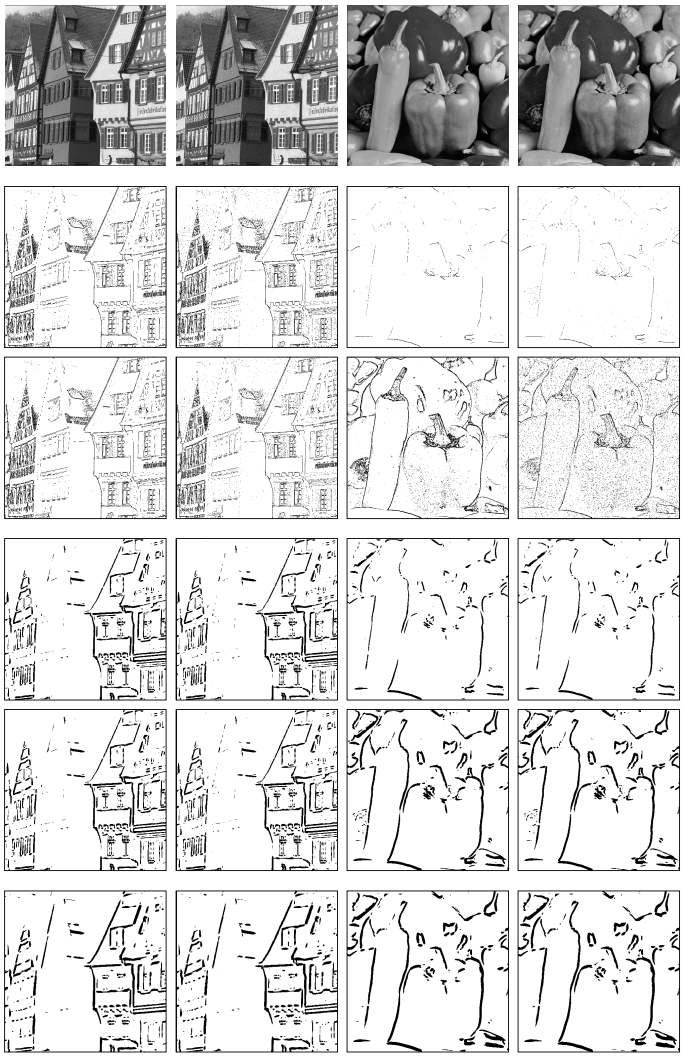


Figure 5.6: Effects of different wavelet thresholding-based techniques for edge detection. Four test images are considered (row 1, from left to right): the distortion-free image "Houses", the same image with added GWN of $\sigma_n = 10$, the distortion-free image "Peppers", the same image with added GWN of $\sigma_n = 10$. The corresponding edge maps are obtained by wavelet thresholding according to five different strategies (rows 2 to 6, from top to bottom): at scale 2^1, selecting wavelet coefficients whose magnitude is above a fixed threshold value; at scale 2^1, choosing 5% of the highest wavelet coefficient magnitudes; the same two methods but now applied on the wavelet coefficients at scale 2^3; and the proposed method of choosing 5% of the highest wavelet inter-scale products $\mathrm{P}_{3\to4}$.

ucts [Zhang and Bao, 2002,Bao et al., 2005], we opt for the following content adaptive thresholding procedure: the edge map is determined by the locations of $\mu \in [0, 100]\%$ largest coefficients of inter-scale product. Hereafter, the percentage μ is referred to as the *threshold percent index* and the corresponding threshold value is denoted T_μ.

For completeness, row 6 of Figure 5.6 shows the results of the proposed scale product-based method applied on the images of "Houses" and "Peppers", noise-free and noisy; in particular, we use $\mathrm{P}_{3\to4}$ and $\mu = 5\%$. The two main points of improvement over the same thresholding technique applied on scale 2^3 (row 5 of Figure 5.6) are the following: for both image contents, there is now less tiny structures in the edges maps, and for the "Houses" image, the darkest house is now better represented (note the outline of the roof which was nearly undetected from a single scale 2^3 but is nicely delineated from the scale product $\mathrm{P}_{3\to4}$). Alternatively, applying a fixed instead of an adaptive threshold on inter-scale product results in similar improvements (details not shown) but the problem of detecting fewer visually significant edges in "Peppers" image stays.

Next, we examine in more detail the effects of image degradations on the proposed edge detection scheme. Here, we study the effects of blur and noise independently. The effects of multiple distortions (blurred images with added noise) are examined later in Section 5.8.2. Figure 5.7 and Figure 5.8 show the edge maps detected by our proposed method for the image "Plane" from the LIVE database (described in Section 5.8.1.2) distorted, respectively, by GBlur and by additive GWN. We consider five different versions of the image: the undistorted image referred to as the REF and two blurry versions (all taken from the LIVE database), and two noisy versions (created by adding GWN to the REF according to the process described in Section 5.2).[7] Rows 2 to 4 of each Figure 5.7 and Figure 5.8 represent results obtained with different combinations of wavelet scales used for the inter-scale products, specifically (from top to bottom): $\mathrm{P}_{2\to3}$, $\mathrm{P}_{2\to4}$ and $\mathrm{P}_{3\to4}$.

In Figure 5.7, we demonstrate that the proposed wavelet-based edge detection mechanism is little sensitive to blur, less so when the coarser scales are used for the computations. This behavior is expected since we include in the edge map only the the strongest edges, and those are least affected by blur. Yet a bigger challenge for an edge detection scheme is the case of images contaminated by noise. As can be observed from Figure 5.8, the proposed technique is able to successfully respond to the challenge. It is important to notice that, while in the case of noise-free images the particular choice of working wavelet scales resulted in minor differences between the edge maps, this particular choice appears crucial in the case of very noisy images (GWN of $\sigma_{n2} = 25$). Compare, for example, the far left and the far right edge maps in Figure 5.7. We clearly observe that the maps from $\mathrm{P}_{3\to4}$ remain nearly unchanged by the addition of noise while the maps from $\mathrm{P}_{2\to3}$ get "contaminated" by pixels which are in fact not true edges but "artifacts" of the noise (see especially the area of grass

[7] For the same REF and blurry images of "Plane", [Liu and Heynderickx, 2011] demonstrated the deteriorating effects of blur on the results of Sobel detector; we refer to Figure 5.25 for illustration.

in the bottom of the image).

5.3.3 ACR estimate of the local Lipschitz exponent

As outlined in the introduction of this chapter, digital image blur is often coupled to the edges in the image. Visually, edges are observed as sharp (abrupt) transitions, or "discontinuities", in image pixel values. In mathematics as well as in image processing, edges are observed as *singularities* (non-differentiable points) and as such can be characterized by Lipschitz exponents. By definition, a function $f(t)$ is *Lipschitz* α over an interval $[a, b]$ if for all $t \in [a, b]$ there exists a constant C such that

$$\forall t \in [a, b], |f(t) - f(t_0)| \leq C|t - t_0|^{\alpha}. \tag{5.13}$$

The *Lipschitz regularity* of $f(t)$ at t_0 over $[a, b]$ is the superior bound of all α satisfying Eq.(5.13). If a function $f(t)$ has a *singularity* at point $t = t_0$ (*i.e.* $f(t)$ is not differentiable at t_0) then the Lipschitz exponent $\alpha < 1$ at t_0 describes this singular behavior [Mallat, 1999]. Figure 5.9 illustrates the first five scales of the DWT applied on the typical types of edges described earlier in Section 5.7 and illustrated in Figure5.3. We consider both the noise-free (left plots in the figure) and the noisy signals (right plots in the figure). Note for a Dirac structure edge the fast decrease of wavelet amplitudes over the scales ($\alpha = -1$). On the contrary, the coefficients of a step and a roof structure edges tend to increase or keep invariant over the increasing scale ($\alpha \geq 0$). As discussed in the previous sections, the detail coefficients of noise disappear quickly with the increasing DWT scale.

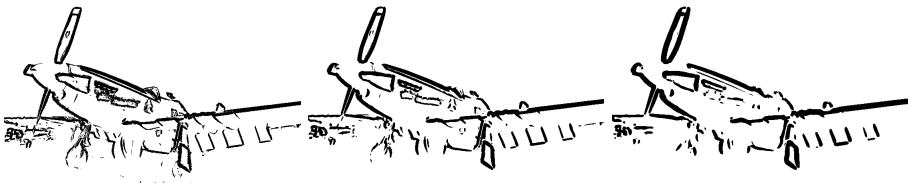
Details about mathematical characterization of singularities using α can be found in the works of [Jaffard, 1991, Mallat and Hwang, 1992, Mallat and Zhong, 1992, Malfait and Roose, 1997, Hsung et al., 1999] who proposed early techniques for estimation of α using multiscale image representation. They showed that the maxima of the wavelet transform modulus can detect the locations of the irregular structures and proposed a numerical procedure to estimate local Lipschitz exponents of these irregularities. However, the early approaches were computationally demanding [Mallat and Zhong, 1992] or imprecise [Malfait and Roose, 1997]. Thus, further research was needed to allow for wavelet-based edge characterization to be used in practical applications.

One such advancement that, next to being computationally efficient, has a major advantage of being highly robust to noise has been proposed by [Pižurica et al., 2002]. In fact, the authors formulate two wavelet-based measures of local regularity: the *average point ratio* (APR) and the *average cone ratio* (ACR). In essence, both quantities are inter-scale ratios of wavelet coefficients but they differ in that the APR tracks the evolution of the *"individual"* wavelet coefficients while the ACR describes the *"collective"* evolution of the wavelet coefficients; the exact definitions are given next.

Let 2^n and 2^k denote any two dyadic scales such that $n, k \in \mathcal{Z}$ and $k \geq n + 1$; $|w_{j,l}|$ is a magnitude of a wavelet coefficient at scale j, $j \in [n, k]$ at position l and

REF | REF + GBlur (σ_1=1.708) | REF + GBlur (σ_2=4.917)

$E_{2\to3}$: Edge maps from wavelet scales 2^2 and 2^3

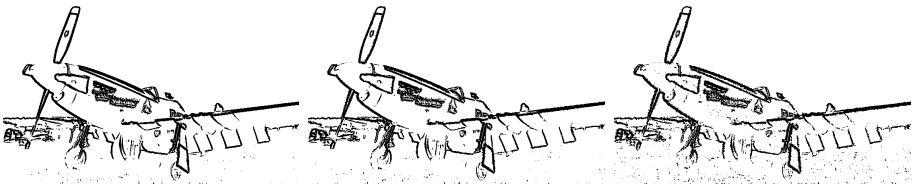


$E_{2\to4}$: Edge maps from wavelet scales 2^2, 2^3 and 2^4

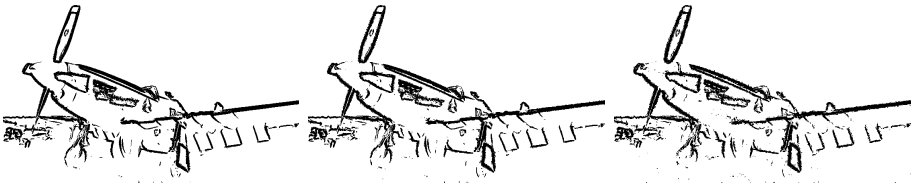
$E_{3\to4}$: Edge maps from wavelet scales 2^3 and 2^4

Figure 5.7: Effects of GBlur on the proposed edge detection scheme. Edge maps obtained by thresholding of wavelet inter-scale products (from top to bottom): $\mathrm{P}_{2\to3}$, $\mathrm{P}_{2\to4}$ and $\mathrm{P}_{3\to4}$.

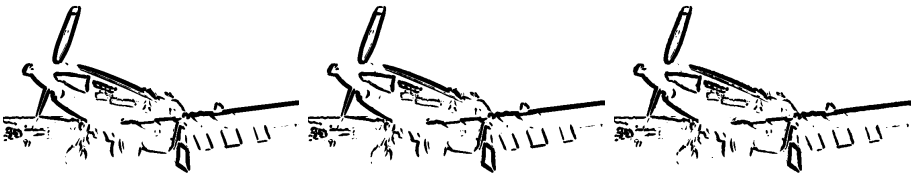
REF REF + GWN (σ_{n1}=10) REF + GWN (σ_{n2}=25)



$E_{2\ 3}$: Edge maps from wavelet scales 2^2 and 2^3



$E_{2\ 4}$: Edge maps from wavelet scales 2^2, 2^3 and 2^4



$E_{3\ 4}$: Edge maps from wavelet scales 2^3 and 2^4

Figure 5.8: Effects of GWN on the proposed edge detection scheme. Edge maps obtained by thresholding of wavelet inter-scale products (from top to bottom): $P_{2\to3}$, $P_{2\to4}$ and $P_{3\to4}$.

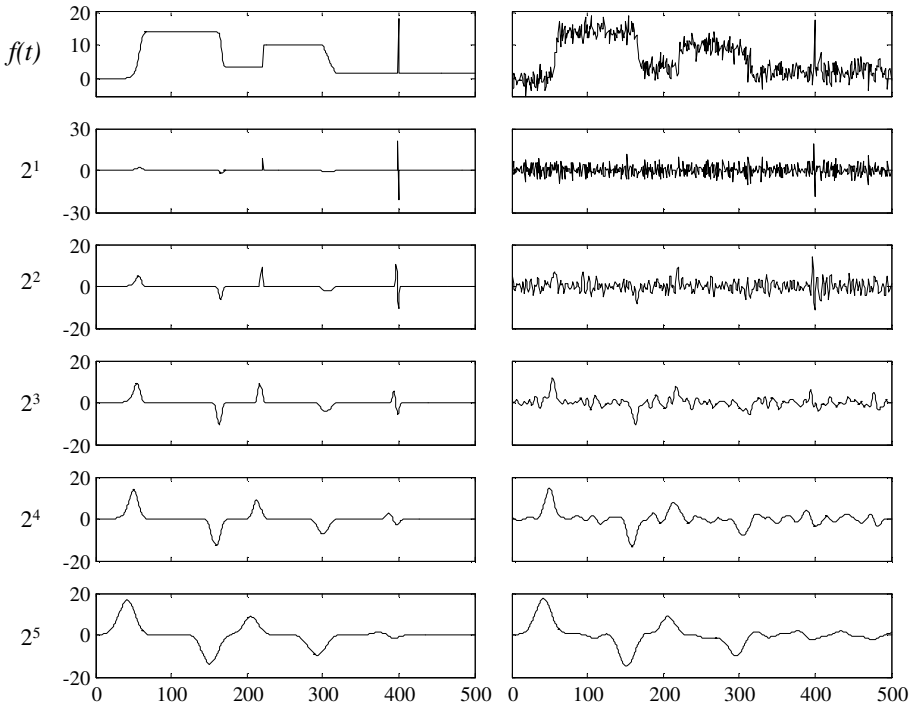


Figure 5.9: An illustration of the evolution of the wavelet coefficients across the scales for a noise-free signal $f(t)$ (left) and for its noisy version (right). Note how the coefficients corresponding to noise diminsh quickly and those corresponding to signal discontinuities survive across the scales.

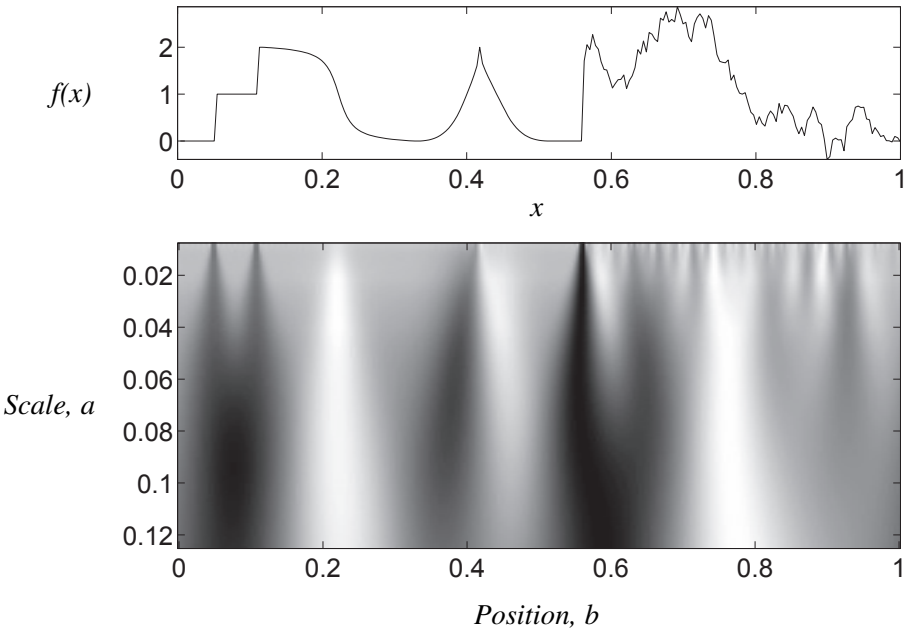


Figure 5.10: [Adapted from [Mallat, 1999]] For a given position b and a given (normalized continuous) scale a, the cone of influence determines the set of wavelet coefficients influenced by the value of the signal at the specified position: (top) signal $f(t)$ and (bottom) wavelet transform of $f(t)$. Black, gray and white points correspond respectively to positive, zero and negative wavelet coefficients.

$l \in \{1, \ldots, L\}$. Then, the APR is defined as

$$\alpha_{n \to k,l} \triangleq \log_2 \left(\frac{1}{k-n} \sum_{j=n}^{k-1} \frac{|w_{j+1,l}|}{|w_{j,l}|} \right). \tag{5.14}$$

and it acts as a rough estimate of the local Lipschitz exponent α.

Next, in the context of the ACR measure, we briefly introduce the concept of a *cone of influence*; for a more detailed overview the reader is referred to [Mallat, 1999]. Figure 5.10 illustrates the behavior of wavelet coefficients across scales, both near and away from singularities of the input signal $f(t)$. We clearly notice that singularities create large amplitude coefficients not only at their own spatial position but also within a certain neighborhood region where that region becomes larger at larger scales. In general, for singularities as well as for non-singularities, for a given position l and a given scale j, *the cone of influence* $C(j,l)$ determines the set of positions at which wavelet coefficients are influenced by the value of the signal at the specified position.

Now, if we denote by $I_{j,l}$ the sum of a discrete set of the magnitudes of the wavelet coefficients at the resolution scale 2^j which belong to the cone of influence of the

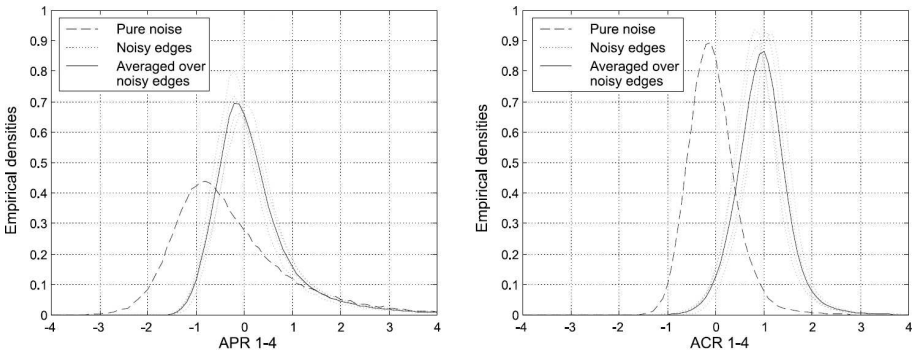


Figure 5.11: [Taken from [Pižurica et al., 2002]] Conditional densities of (a) APR and (b) ACR, computed from scales 2^1 - 2^4. The standard deviation of added noise is $\sigma_{n2} = 25$.

analyzed position l, the ACR quantity can be defined as

$$\beta_{n\to k,l} \triangleq \log_2 \left(\frac{1}{k-n} \sum_{j=n}^{k-1} \frac{I_{j+1,l}}{I_{j,l}} \right), \quad I_{j,l} \triangleq \sum_{m \in C(j,l)} |w_{j,m}|. \tag{5.15}$$

In general, calculating ratios of (sums of) wavelet coefficients such as those in Eq.(5.14) and Eq.(5.15) can result in numerical instabilities (division by very small numbers). This was not an issue for our experiments reported here as we compute the ACR values only at the positions of image edges where the wavelet coefficients (and especially the sums of the coefficient magnitudes) are not close to zero ($I_{j,l} \gg 0$). Otherwise, to keep the practical implementation of the methods general and prevent potential problems, we ought to consider only those ratios for which the denominator amplitude is above a certain very small threshold value.

According to Eq.(5.15), the ACR can be seen as a measure of evolution of the wavelet coefficients across adjacent dyadic scales and inside a cone of influence centered at a given spatial position. As for the APR, [Pižurica et al., 2002] have shown that ACR is a good estimate of the local Lipschitz exponent, in particular $\alpha + 1$. Moreover, the study demonstrated that the ACR measure is able to successfully separate the noise from the useful edges, notably better than the APR. The major advantage of ACR over APR is in its superior immunity to noise even at very high levels of noise such as GWN of $\sigma_{n2} = 25$ is illustrated in Figure 5.11. The ability of ACR to characterize edges (estimate local edge regularity) while being nearly insensitive to noise is further explored in the following Section 5.4 where we introduce the novel ACR-based measure of image blurriness.

Moreover, what is of particular interest for our application, the ACR measure is sensitive to image blur. This property is illustrated by the PDFs of ACR2-4 shown in

Figure 5.12. There, we consider four variants of the well-known "Lena" image – the distortion-free image, the image with added GWN, the image with added DBlur, and the image with added both DBlur and GWN (the distortion parameters are $r = 3$ for DBlur and $\sigma_n = 25$ for GWN). It is obvious from the plot of ACR distributions that the measure reacts to the changes of blur in the image but not to the changes of noise. This property of the ACR measure is exploited in the next section where we introduce the novel ACR-based measure of image blurriness.

5.4 New ACR-based noise immune NR measure of blurriness: CogACR

Our proposed algorithm for estimating the level of image blurriness is illustrated in Figure 5.13. Once in the wavelet domain, we first aim to select the set of spatial positions in the image $\mathbf{x}$ for which to calculate the ACR measure – the edge map $\mathbf{e}$ for the image. This is done following the procedure described in Section 5.3.2. Thus, the first three control parameters of the method are: the percent μ of image pixels which are allowed in the edge map (threshold parameter), and the two identifiers of boundary scales n_1 and k_1 for the inter-scale product of wavelet coefficients $\mathrm{P}_{n_1 \to k_1}$.

In the next step, we use Eq. (5.15) to calculate ACR values $\beta_{n_2 \to k_2, l}$ for all edge positions $l = 1, \ldots, L$ from the edge map $\mathbf{e}$. Computing ACR quantities requires another two control parameters to be chosen: the lower and the upper boundary scale for the ACR computations, *i.e.*, the parameters n_2 and k_2. Subsequently, we construct a histogram of ACR values which we call "HistACR" and denote by $\mathbf{h}$. For convenience of notation, we introduce a function $\mathrm{hacr}((x))$ to represent the described process of computing the histogram of ACR coefficients for image (x), thus $\mathbf{h} = \mathrm{hacr}((x))$. The elements of this histogram $h_b, b = 1, \ldots, B$ can be described as

$$h_b = \eta \sum_{l=1}^{L} \delta_{lb}, \quad \delta_{lb} = \begin{cases} 1, & \text{if } \beta_l \text{ belongs to bin } b \\ 0, & \text{otherwise} \end{cases}, \tag{5.16}$$

where η is chosen such that $\sum_{b=1}^{B} h_b = 1$ and B is the number of histogram bins. Finally, the center of gravity of the ACR histogram $\mathbf{h}$ is computed to quantify the level of blur in the image. We refer to this measure as the CogACR and write

$$\mathrm{cog}(\mathbf{h}) = \frac{\sum_{b=1}^{B} h_b \, \beta_b}{\sum_{b=1}^{B} \beta_b}. \tag{5.17}$$

Let us now take a more detailed look at the control parameters of the proposed CogACR scheme for blur identification. Commonly, we would use the same set of

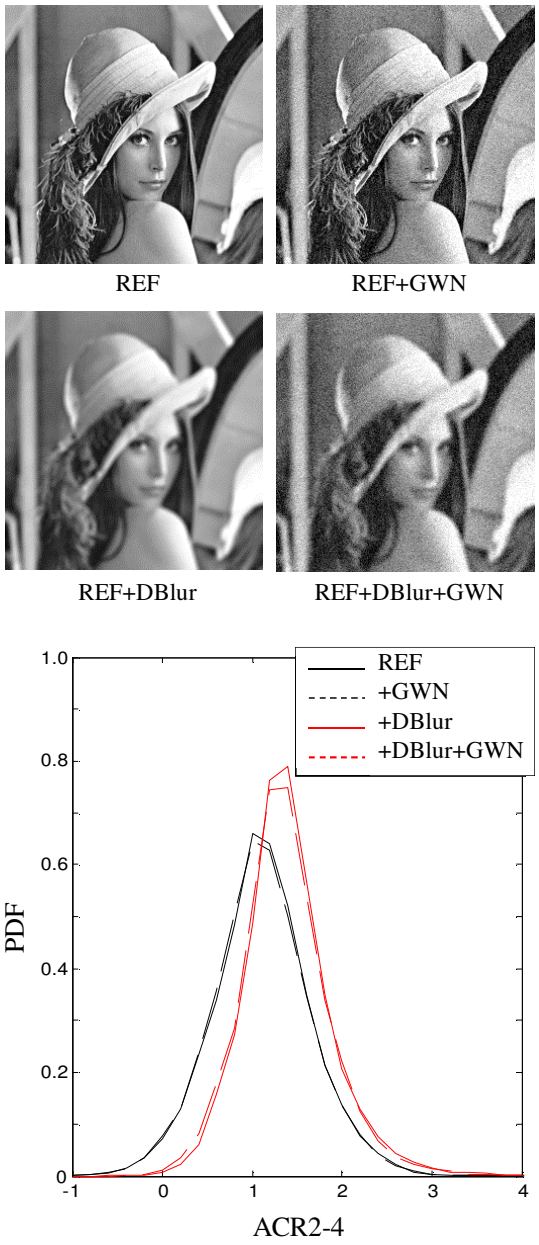


Figure 5.12: The ACR2-4 measure is computed for for variants of the "Lena" image (from left to right, top to bottom) distortion-free, with added GWN of $\sigma_n = 25$, with added DBlur of $r = 3$ pixels, and with added both DBlur of $r = 3$ and GWN of $\sigma_n = 25$. The plot in the bottom shows the PDF of ACR2-4.

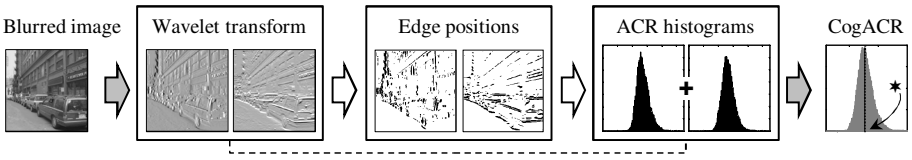


Figure 5.13: Flowchart of the proposed CogACR image blur measure in a no-reference (NR) scenario (the reference image is not available).

wavelet scales for the purpose of edge detection and ACR computation; thus $n = n_1 = n_2$ and $k = k_1 = k_2$. This leaves us with a total of three control parameters of the method: μ, n and k. We first look into the scale identifiers n and k.

Figure 5.14 depicts HistACRs for the three considered parameter-value pairs, $n - k$: ACR2-3, ACR2-4 and ACR3-4. We observe the effects of these parameters for all three types of blur (GBlur, DBlur, MBlur), each at two different levels and each with and without noise ($\sigma_{n1} = 10$ and $\sigma_{n2} = 25$). We notice from the plots that the histograms of each ACR2-3, ACR2-4 and ACR3-4 clearly respond to the increase in the level of blur (when BL increases from zero through BL1 to BL2, the histograms shift to the right which corresponds to the increasing values of the Lipschitz exponents of edge pixels), while at the same time they remain nearly unaffected by the addition of GWN. The responsiveness to blur seems most for ACR2-4, while the immunity to noise appears highest for ACR3-4. Therefore, depending on the specific priority requirement for the blur identification scheme, we would choose ACR2-4 to achieve *high accuracy* in blur estimation and ACR3-4 to achieve *high noise immunity* of the measure. Figures 5.15 and 5.15 show a close-up of the HistACR peaks for ACR2-4 and ACR3-4, respectively.

Next, we examine the effects of the percent μ of image pixels that get into the edge map. Figure 5.17, Figure 5.18, and Figure 5.19 demonstrate the effects of different μ-values on the edge maps and also on the CogACR performance for the scale parameters $n = 2$ and $k = 4$. We notice that these effects are different for different image contents. Overall in terms of CogACR, allowing too many pixels to the edge map (*e.g.* $\mu = 10$ or $\mu = 20$) has a deteriorating effect on the proposed blur identification scheme. This is more pronounced for images such as "Peppers", which contain fewer (strong) edges compared to the "FishingBoat" or the "Houses". Small amount of strong edge pixels allows more noise pixels into the edge map, and "contaminates" the edge map. Consequently, the CogACR robustness to noise decreases (more so at higher levels of blur) and its sensitivity to blur decreases. Conversely, allowing too few pixels to the edge map (*e.g.* $\mu = 1\%$) fails to capture the main adge structure of the image. Therefore, we opt for $\mu = 5\%$ which seems able to capture the main edges in the image and at the same time stays rather unaffected by image noise. Unless otherwise explicitly mentioned, the value of $\mu = 5\%$ is kept the same for all experiments reported in this chapter.

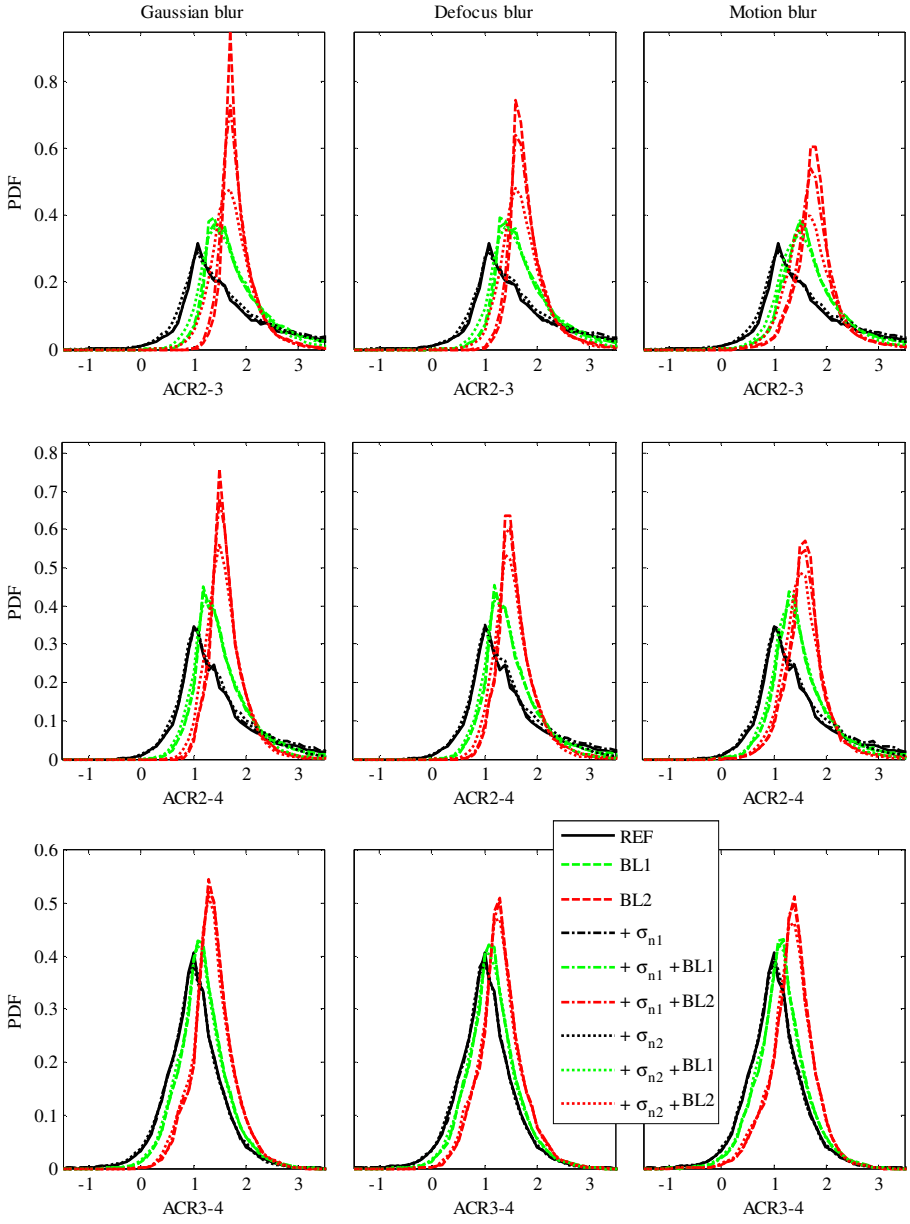


Figure 5.14: PDF of ACR calculated across different scales: ACR2-3, ACR2-4, ACR3-4. The amounts of GBlur, MBlur, DBlur are chosen to correspond to approximately the same PSNR values: 27 dB for BL1 and 24 dB for BL2. The level of added GWN is determined by $\sigma_{n1} = 10$ and $\sigma_{n2} = 25$.

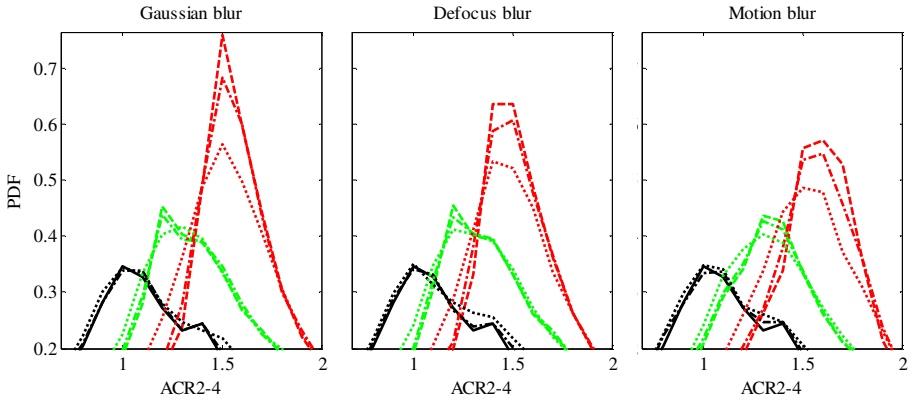


Figure 5.15: Zoom-in on the upper part of the plots for ACR2-4 from Figure 5.14.

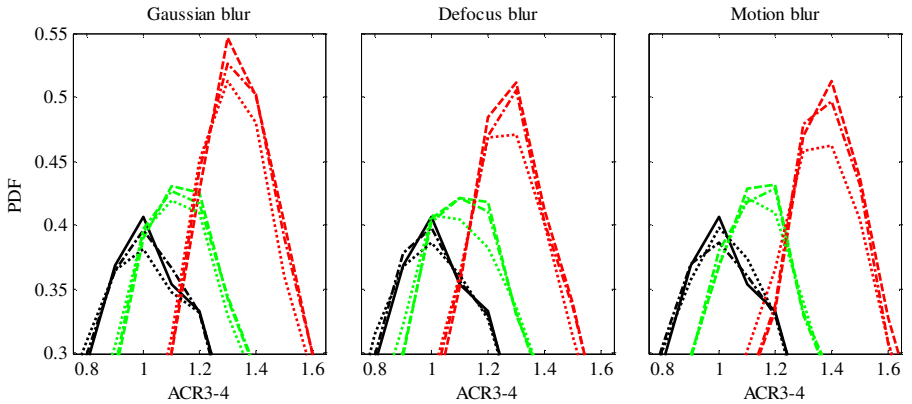


Figure 5.16: Zoom-in on the upper part of the plots for ACR3-4 from Figure 5.14.

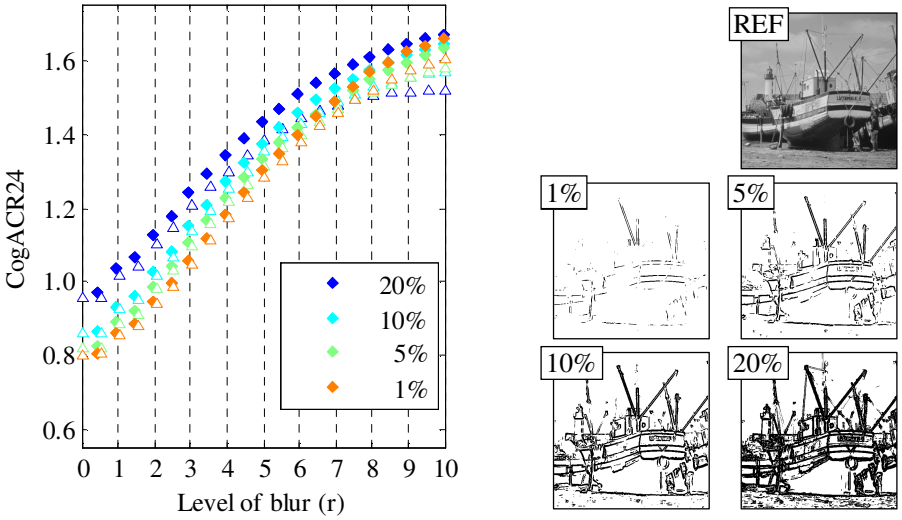


Figure 5.17: Effects of mask selection (edge detection) on CogACR24 measure. (Left) CogACR24 over a range of DBlur, $r = \{0, 0.25, 0.5, \ldots, 10\}$. Diamond marks correspond to noise-free images and triangle marks refer to images with added GWN of $\sigma = 10$. Different colors denote different values of the threshold percent index $\mu = \{1, 5, 10, 20\}\%$. (Right) The REF image "FishingBoat" together with its mask images corresponding to each $\mu = \{1, 5, 10, 20\}\%$.

Next, we examine effects of different μ values on the edge maps for different image contents can be seen in Figure 5.17, Figure 5.18, and Figure 5.19. Our experimental study explored the following search space for the percent coefficient $\mu = \{1, 5, 10, 20\}$. As the value of μ increases, the absolute value of the threshold T_μ will decrease to allow more wavelet coefficients to be included in the mask. For example, in our study we use images of 256×256 pixels. In the case where the threshold percent index is selected to be $\mu = 5$, the number of peak wavelet coefficients to be included in the mask is calculated as $(256 \times 256) \times 5\% = 3276.8$ which is rounded off to $T_5 = 3277$.

5.5 New edge descriptor for edge-based image matching: HistACR

Figure 5.20 illustrates HistACR and CogACR for the images of "Face" and "Cactus". The scores are shown for the blur-free images as well as for three different levels of DBlur, $r = \{2, 5, 10\}$ pixels. As with Figure 5.14, we notice in Figure 5.20(a),(b) that the center of gravity of ACR histograms shifts with the increase of blur in the image which allows the CogACR measure to clearly distinguish between a wide range of blurriness in the image, as we can see in the graph (c) of Figure 5.20. Nevertheless,

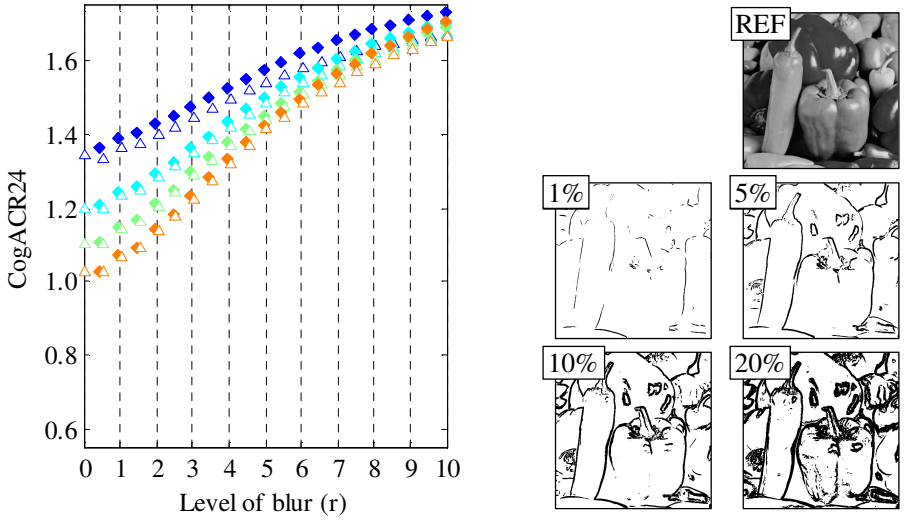


Figure 5.18: The same as Figure 5.17 but for the REF image "Peppers."

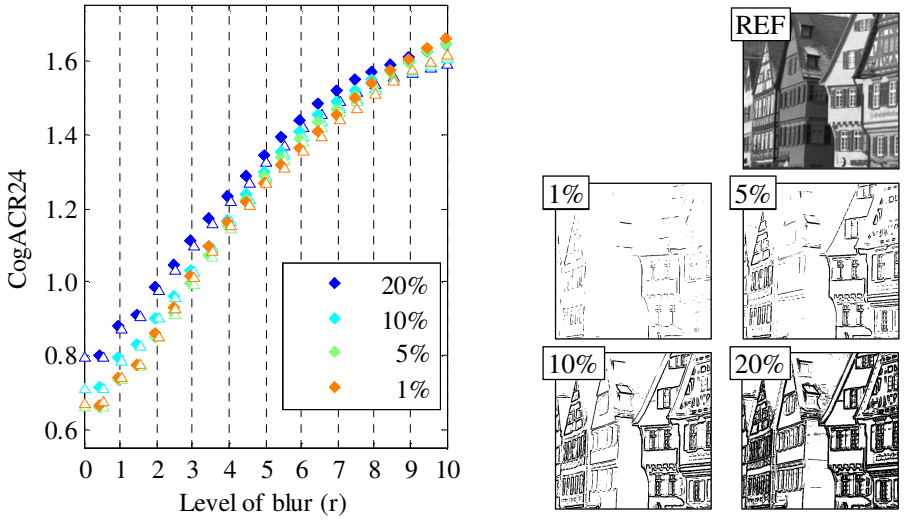


Figure 5.19: The same as Figure 5.17 but for the REF image "Houses."

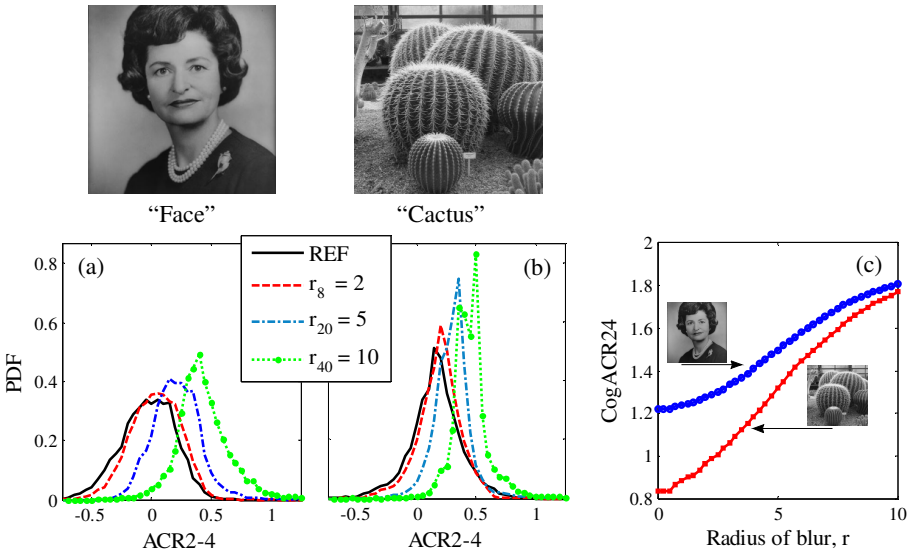


Figure 5.20: (a) Histograms of ACR2-4 coefficients for "Cactus" image; (b) histograms of ACR2-4 coefficients for "Face" image. The plots in (a) and (b) are shown for the REF images as well as for their distorted versions at 3 different levels of DBlur, $r = \{2, 5, 10\}$ pixels. (c) CogACR24 curves for "Cactus" and "Face" images over the same range of DBlur distortions.

we notice also that, while both monotonically rising with the increase of blur, the exact CogACR values at a given blur r differ for the two contents. Moreover, the HistACRs corresponding to the two contents are very different. This is expected behavior and it reflects the fact that the prevailing type and characteristics of edges differ between the two contents.

In Section 5.7 we give more details about the related considerations by [Tong et al., 2004] concerning the influence of blur on different types of edges. In the example from Figure 5.20, the "Cactus" image has more high frequency content compared to the "Lady" image which corresponds to fewer strong step edges in the "Cactus" image and more Dirac edge structures. This observation is reflected in the HistACR as well as in the trends of CogACR measure depicted in Figure 5.20(c). There, at the lower levels of blur ($r = 1$ to $r \approx 6$), the CogACR is more sensitive to the changes of blur in the "Cactus" image than to those in the "Lady" image. As the level of blur increases, the difference between the shapes of the ACR histograms, as well as the CogACR trends for the two images, gradually get diminished. Thus, as already noted in the previous Section 5.4, the CogACR measure is sensitive to the content of images. This is directly related to the variability in edges on which the measure is computed. Depending on the image content, a fixed percent of the selected edges may include different proportions of strong and sharp versus weak and smooth edges.

Given the observed relationship between edge characteristics and the HistACR, and knowing that ACR values correspond to the Lipschitz exponents of the edge pixels, we propose to use HistACR as an *edge descriptor* for the purpose of image matching on the criterion of similarity in reaction to the presence of blur. Unlike many existing image similarity measures that draw from the contextual information in the image, the novel similarity measure is built around the edge-related information in the image. It is designed to distinguish between images with similar edge characteristics.

To measure the distance between two ACR histograms, $\mathbf{h}_1$ and $\mathbf{h}_2$, we explore three common measures known in the literature: histogram intersection, Kullback-Leibler distance (KL), and symmetrical Kullback-Leibler distance (SKL). The measure which proved best suited to the task is the SKL dissimilarity measure defined as

$$d_{\mathrm{SKL}}(\mathbf{h}_1, \mathbf{h}_2) = \frac{1}{2} \sum_{b=1}^{N_{\mathrm{bin}}} (h_{1b} - h_{2b}) \log \left(\frac{h_{1b}}{h_{2b}} \right), \tag{5.18}$$

where h_b is the number of elements in the bin $b = 1, \ldots, N_{\mathrm{bin}}$ of the histogram $\mathbf{h}$.

5.6 HistACR-based image dictionary matching for NR blur identification

As we will demonstrate in the results section, the CogACR measure is able to successfully identify image blur in a no-reference IQA scenario. Nevertheless, as discussed in Section 5.4, the CogACR measure (as well as other blur measures in the literature) is sensitive to image content. In the cases where we expect large diversity of image content, we propose using HistACR as content descriptor and identifying blur as proposed next.

We assume two non-overlapping data sets: a set of training images for which the level fo blur BL is exactly known, $\mathcal{X} = \{\mathbf{x}_i : i = 1, 2, \ldots, N_{\mathrm{tr}}\}$, and a set of testing images for which BL is unknown, $\mathcal{Y} = \{\mathbf{y}_j : j = 1, 2, \ldots, N_{\mathrm{ts}}\}$. Here, N_{tr} and N_{ts} are used to denote the total number of images in the training and in the testing set, respectively.[8]

For a given test image $\mathbf{y}_j$, we are interested to find a training image $\mathbf{x}_i$ which is most similar to $\mathbf{y}_j$ in terms of edge properties. In that sense, the problem can be seen as the best match problem, or a well-known image dictionary matching problem [Cha, 2000]. In our case, the similarity criteria is twofold: (1) the CogACR values of the two images are similar, and (2) the ACR histograms of the two images are similar. Correspondingly, our algorithm for finding the best matching image is a two-stage process, including a *candidate selection* and a *candidate verification* stage. A

[8]Note that this is different from Chapter 3 and Chapter 4 where we use N_{tr} and N_{ts} to denote the total number of *pairs* of respectively training and testing images. The details are explained in the related chapters.

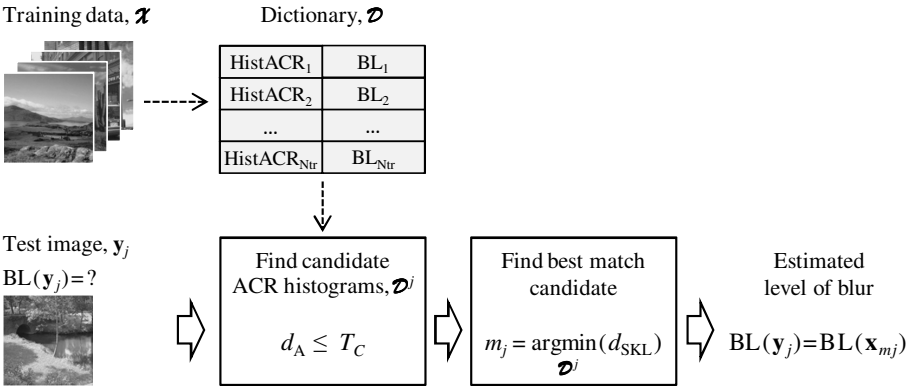


Figure 5.21: Flowchart of the proposed technique for dictionary-based NR image blur identification using the CogACR measure.

schematic illustration of the method is presented in Figure 5.21. We explain the details shortly.

5.6.1 Candidate selection

As illustrated in Figure 5.21, the input and output of the proposed method for automated NR blur identification are:

Input: A test image $\mathbf{y}_j$, a dictionary of ACR histograms $\mathcal{D}$.

Output: Estimated blur level $\mathrm{BL}(\mathbf{y}_j)$.

First, the training dataset $\mathcal{X}$ is used to build a *dictionary* $\mathcal{D}$. Each entry of the dictionary corresponds to one training image $\mathbf{x}_i$, $i = 1, \ldots, N_{\mathrm{tr}}$. The dictionary entries are ordered pairs $(\mathbf{h}_{\mathbf{x}i}, \mathrm{BL}_{xi})$ in which the first element is the histogram of ACR coefficients $\mathbf{h}_{\mathbf{x}i} = \mathrm{hacr}(\mathbf{x}_i)$, and the second element is the corresponding known level of blur, BL_{xi}. Thus, the dictionary can be described as

$$\mathcal{D} = \{(\mathbf{h}_{\mathbf{x}i}, \mathrm{BL}_{xi}) \ : \ \mathbf{h}_{\mathbf{x}i} = \mathrm{hacr}(\mathbf{x}_i),\ i = 1, \ldots, N_{\mathrm{tr}}\} \tag{5.19}$$

After the dictionary has been created, we compute the histogram of ACR coefficients also for the test images $\mathbf{h}_{\mathbf{y}_j} = \mathrm{hacr}(\mathbf{y}_j)$, $j = 1, \ldots, N_{\mathrm{ts}}$. Now, for each $\mathbf{h}_{\mathbf{y}_j}$ we search the dictionary for all entries that satisfy the first similarity criterion: the ACR histograms of the images have similar CogACR values. We use the absolute difference as a similarity measure between CogACR values:

$$d_{\mathrm{A}}\left(\mathrm{cog}\left(\mathbf{h}_{\mathbf{y}_j}\right), \mathrm{cog}\left(\mathbf{h}_{\mathbf{x}i}\right)\right) = \left|\mathrm{cog}\left(\mathbf{h}_{\mathbf{y}_j}\right) - \mathrm{cog}\left(\mathbf{h}_{\mathbf{x}i}\right)\right|, \tag{5.20}$$

where $|\cdot|$ means the absolute value. Then, the subset of the dictionary entries which fall in the T_{C} neighborhood of the $\mathrm{cog}(\mathbf{h}_{\mathbf{y}_j}$ constitutes the *candidate set*, $\mathcal{D}^j \subset \mathcal{D}$:

$$\mathcal{D}^j = \left\{ (\mathbf{h}_{\mathbf{x}q}, \mathrm{BL}_{xq}) : \, d_{\mathrm{A}} \left(\mathrm{cog} \left(\mathbf{h}_{\mathbf{y}_j} \right), \mathrm{cog} \left(\mathbf{h}_{\mathbf{x}q} \right) \right) \leq T_{\mathrm{C}}, \, q = 1, \ldots, N_C \right\}. \tag{5.21}$$

Clearly, the smaller the threshold value of T_{C} the smaller the number of candidates N_C where $N_C \leq N_{\mathrm{tr}}$. Accordingly, in view of the scheme in Figure 5.21, the value of the threshold T_{C} can be seen as a configurable parameter of the system. The influence of T_{C} on the classification process is further discussed in the results Section 5.8.4.

As remarked in [Cha, 2000], the candidate selection stage of image dictionary matching problem can be seen as a variation of the k-nearest neighbor search problem. Namely, similar to the k-nearest neighbor problem where the goal is to find k-nearest neighbors in the reference set, the candidate selection problem we described aims to find *all* dictionary entries whose distances are within the neighborhood determined by the threshold T_{C}.

5.6.2 Candidate verification

The goal of the candidate verification stage is to select from the the candidate set $\mathcal{D}^j$ the best matching candidate for the given test image $\mathbf{y}_j$. To do that, we first compute similarities (distances) between ACR histograms of the test image and those of the candidate images. For this purpose, we use the dissimilarity measure d_{SKL} defined by Eq. (5.18). Finally, the blur of the test images is estimated equal to the blur level of the candidate entry with the smallest value of d_{SKL}. If we use $argmin(\cdot)$ to denote the argument of the minimum, the candidate verification process can be described by the following formulation

$$m = \underset{q}{\mathrm{argmin}} \left(d_{\mathrm{SKL}} \left(\mathbf{h}_{\mathbf{y}_j}, \mathbf{h}_{\mathbf{x}q} \right) \right), \;\; \mathbf{h}_{\mathbf{x}q} \in \mathcal{D}^j, \tag{5.22}$$

$$\mathrm{BL}_{y_j} = \mathrm{BL}_{xm}, \;\; \mathrm{BL}_{xm} \in \mathcal{D}^j. \tag{5.23}$$

5.7 Existing NR blur measures

In this section, we briefly review the basic principles and considerations of NR blur measure from the literature. Some additional related reviews can be found in [Firestone et al., 1991, Ferzli and Karam, 2009]. The key characteristics of the methods are also summarized in Table 5.1. The most recent and best performing of these measures are used in the comparative analysis of NR blur measure provided in the results section.

1. **LipschitzCG.** [Rooms et al., 2002] rely on the concept of *Lipschitz exponent* as the descriptor of edge singularities (see Section 5.3.3 for more details). First,

to select pixels which correspond to the sharpest edges in the image, they apply thresholding to the gradient of the wavelet coefficients at the highest resolution scale – all pixels mapped to the gradient above a certain (empirical) threshold are considered edges. Next, relying on the findings of [Mallat and Hwang, 1992, Mallat and Hwang, 1992], they compute the Lipschitz exponent for each edge pixel by fitting an exponential curve to the modulus maxima of the wavelet coefficients. Finally, the center of gravity of the histogram of estimated Lipschitz exponents of all edge pixels in the image is computed to represent the global measure of the level of blur in the image.

2. **EdgeAPR.** For defocus estimation, [Lin et al., 2004] propose exploiting the ratio of the wavelet coefficients at two adjacent scales. This ratio corresponds to the concept of *average point ratio* (APR) from [Pižurica et al., 2002] (see Section 5.3.3 for more details). Specifically, the method is comprised of the following four steps: (1) apply a two-level DWT to the image (*e.g.* Haar transform); (2) at scale 2^1, find horizontal and vertical edges;[9] (3) for every edge pixel, compute the magnitude ratio of wavelet coefficient at scale 2^1 over that at scale 2^2 and average these ratios over all pixels of a given edge; and (4) find the maximum of such ratios over all edges detected in step (2) and take it as a measure of defocus, EdgeAPR. Note that the value of EdgeAPR decreases as the amount of DBlur increases.

3. **EdgeWidth.** Primarily aiming at IQA for encoded digital images, [Marziliano et al., 2004] suggest measuring the spread of edges directly in the spatial domain. They first apply a Sobel filter to detect edges, then compute the width of every detected edge, and finally take the average of all edge widths as a global measure of blur for the image.

4. **EdgeType** [Tong et al., 2004] frame their blur detection scheme around the concept of *edge structure*. Namely, they classify edges into the following four types: A-step structure, Dirac structure, G-Step structure, and Roof structure (see Figure 5.3 for a graphical description). In addition, they assume that most natural images contain all these four types of edges, that most G-step and Roof structures are sharp enough in blur-free images, and that Dirac and A-step structures disappear with blur. Eventually, the authors formulate a number of rules from the properties of Haar DWT and its evolution across scales. They use those formulations to determine: the edge points, the type of an edge structure, and the presence of blur (for G-step and Roof structure). With that established, they

[9] The authors provide no details of the edge detection process. For the purpose of experiments, we implement this step of the algorithm as wavelet thresholding keeping the 5% of the highest wavelet coefficients (the same as in our proposed method). To improve the quality of the edge map, we remove all detected edges comprised of less than 0.1% of a total number of pixels in the image as well as all objects touching the image borders. Finally, to avoid outliers caused by the very small values in the denominator (close to zero), we remove from the edge map all pixels for which the wavelet coefficients at scale 2^2 are smaller than 1.

adopt the following line of reasoning: (1) to judge whether an image is blurred, observe the ratio of the number of pixels that belong to Dirac and A-step structures over all edge pixels – if the ratio is low (below a certain threshold), the image is blurred and vice versa; and (2) to estimate the amount of blur in the image, calculate the percentage of G-step and Roof structures which are blurred (above a certain threshold) – the larger the percentage, the more blurry the image is.

5. **Kurt2-Freq.** The measure proposed by [Zhang et al., 1997] and further detailed by [Zhang et al., 2003, Zhang et al., 2005], builds on the observation by [Vladár et al., 1998] that if one image is visually sharper than the other, then the high spatial frequency components of the first image are larger than those of the second image. To measure the extent of blur in an image, in their case 2D micrographs from a scanning electron microscope (SEM), first the 2D Fourier transform is applied and then a bivariate kurtosis of the 2D power spectral density function is computed, where the spectral density is treated as a probability density function (PDF).

6. **Kurt2-WV.** As [Rooms et al., 2002], [Ferzli et al., 2005] refer to the theory of Lipschitz regularity and the properties of the amplitude of the wavelet transform modulus maxima demonstrated in [Mallat and Hwang, 1992]: it increases with the scale for edge singularities (positive Lipschitz) and it decreases with the scale for noise singularities (negative Lipschitz). The authors propose using the concept of the Kurt2 measure but computed in the wavelet rather than in the frequency domain, thus the bivariate kurtosis of wavelet coefficients. In order to avoid major effects of noise, they suggest working at a wavelet scale high enough to reflect mainly edge content and not the noise; specifically, they chose the discrete dyadic wavelet transform [Zhan and Karam, 2003] at the scale 2^3.

7. **EdgeWidth-WV.** [Ferzli and Karam, 2005] propose using the concept of the EdgeWidth measure but computed in the wavelet instead of in the spatial domain. Their arguments for switching to the wavelet space are the same as with Kurt2-WV and likewise they opt for the same type of the DWT and the same working scale 2^3.

8. **JNBM.** [Ferzli and Karam, 2009] focuse on automatically measuring the extent of blur across images with diverse scene content, the goal which the preceding measures failed to fulfill, as evidenced by their reported test results. Their small-scale experiment with human subjects[10] suggests that humans are able

[10]Each of the 4 participating subjects viewed 6 pairs of Gaussian blurred images to choose the more blurred one from each pair. The pairs were created from the following four image scenes (all included in the LIVE database described in Section 5.8.1.2): Houses, Man, FishingBoat, Peppers. All pairs were different, containing two different scenes each with a different amount of added blur. The sequence of displayed combinations was random.

to differentiate the level of blurriness between different image contents, even when the difference in the amount of blur is small; *i.e.*, even though the content of the two compared images was different, humans were able to perceive the difference in blurriness of the two images.[11] Aiming to design a perceptually relevant measure of blurriness, [Ferzli and Karam, 2009] explored the concept of *just noticeable blur* (JNB) – the term used to denote the minimum amount of blur around an edge which can be detected by the HVS given a contrast higher than the just noticeable difference (JND). To determine the relationship between the local edge contrast and the JNB, the authors conducted another larger-scale human experiment[12] and measured the width of an edge corresponding to the JNB to be $a_{\mathrm{JNB}} = 5$ pixels for the local contrast upto 50 and $a_{\mathrm{JNB}} = 3$ pixels for higher contrasts. Here, the width of edge was measured based on [Marziliano et al., 2004]. The following algorithm for quantifying perceived blurriness is proposed: (1) perform Sobel edge detection, (2) select all 64×64 image blocks which have more than a certain number of edge pixels (edge blocks), (3) estimate the local contrast for each edge block and map it to the corresponding value of a_{JNB}, (4) compute the width for each edge from an edge block a_{i}, (5) estimate the block distortion as a scaled sum of absolute ratios of a_{i} over a_{JNB}, and (6) compute the overall distortion by pooling block distortions.

9. **CPBD.** Another measure that makes use of the JNB concept is named cumulative probability of blur detection (CPBD) [Narvekar and Karam, 2009,Narvekar and Karam, 2011]. The authors start from the individual probabilities of detecting blur for each edge and devise a new framework to combine the individual scores into one cumulative measure. In fact, the CPBD differs from the JNB measure in steps (5) and (6) of the algorithm: the scaled sum of the individual distortions from the JNB method is now replaced by the percentage of edges at which blur is not likely to be detected (in the JNB sense).

10. **LPC.** [Hassen et al., 2010] identify image blur as a strong local phase coherence evaluated in the complex wavelet transform domain. Unlike most of the other methods, this one is not specifically tuned to image blurriness, rather, it is able to detect also other image distortions whics may affect perceptual "sharpness", such as compression, median filtering, and noise contamination.

[11] Note that, in general, the difference in the extent of blurriness (or other type of IQ distortion) is more straight forward to notice when comparing images of the same scene rather than comparing images of different scene content. For further remarks and discussion around the effects of image content in blur identification, the reader is referred to Section 5.8.4.

[12] The images were created as a flat background with a flat square as foreground. To change the contrast, intensity values of the background and foreground were changed within a discrete range of grayscale values, such that the contrast is always greater than the JND. The experiment involved 18 subjects, each viewing 27 different contrasts at one of the 6 tested standard deviations of GBlur. For each presented image, the subject was asked to indicate whether they detected blur or not.

11. **FISH.** The "Fast Image Sharpness" (FISH) method is a wavelet-based technique proposed by [Vu and Chandler, 2012] and targeted at estimating global as well as local image sharpness. The image is first transformed by a three-level separable DWT and then the log-energies of the wavelet subbands are computed. Last, a weighted average of these log-energies is taken as a measure of the overall image sharpness.

12. **S3.** The method proposed by [Vu et al., 2012] exploits both spectral and spatial properties of the image. They measure, per image block: (1) the slope of the magnitude spectrum and (2) the total spatial variation adjusted to account for visual perception (nonlinear visual summation across space). These two quantities are combined into a summary image measure using a weighted geometric mean.

13. **SpaQF.** The method proposed by [Soleimani et al., 2013] is yet another wavelet-based algorithm which relies on Lipschitz regularity of the signals. Following the approach of [Ducottet et al., 2004], the SpaQF method considers contours as smoothed singularities of three particular types: transition, peak, and line. Then, the prediction of the level of blur in the image is based on the comparison between the maxima function extracted from a particular edge point and the theoretical maxima functions of the three edge models. The measure is targeted specifically at GBlur.

5.8 Experimental results

The performance of the proposed techniques is extensively tested and evaluated with respect to a range of parameters, including image content (three public databases are considered), level and type of image blur (Gaussian, defocus, and motion), as well as the presence of varying levels of image noise. The results are organized as follows.

We first describe in Section 5.8.1 the three public databases used in the experiments. Next, in Section 5.8.2, we evaluate the performance of the proposed algorithm for edge detection. Then, in Section 5.8.3, we report on the experiments in which the new HistACR descriptor is used for assessing edge-based image similarity (similar images react similarly to the presence blur).

After testing the methods for edge detection and edge-based content selection, we move to the evaluation of the proposed CogACR method for assessing image blurriness. In these experiments, the proposed methods are evaluated against ten state-of-the-art NR blur measures reviewed in Section 5.7. First, in Section 5.8.4, we perform a series of experiments using CogACR for estimating the radius of DBlur. These experiments involve a large variety of image content and hence we employ the blur estimation technique from Section 5.6 which involves HistACR-based image dictionary matching. Alternatively, in Section 5.8.5, we evaluate performance of the CogACR

Table 5.1: Comparison of the selected NR measures of image blurriness

Measure	Application	Method	Working domain	Edge detection	Content dependency	Noise dependency	HVS aware
LipschitzCG	image restoration, autofocus	magnitude of the Lipschitz exponent for the sharpest edges	wavelet	wavelet thresholding	–	–	no
EdgeAPR	autofocus (tracking objects)	ratio of WV coefficients at 2 adjacent scales	wavelet	no details	–	minor	no
EdgeWidth	source coding optimization, network resource management	width of edges	spatial	Sobel	*e.g.* varying depth of field	–	yes
EdgeType	digital cameras	structure and sharpness of the edges deduced from Haar DWT	wavelet	wavelet thresholding	–	assume no noise	no
Kurt2-Freq	fully/semi-automated SEM systems	bivariate kurtosis of 2D power spectral density	Fourier	no	–	–	no
EdgeWidth-WV	general	width of edges computed at scale 2^3 of DWT	wavelet	Sobel	–	at very high noise	yes
Kurt2-WV	digital video, automated SEM systems	bivariate kurtosis of DWT coefficients at scale 2^3	wavelet	no	yes	not at low/mid SNR	no
JNB	general perceived IQA, image compression	ratio of edge- and JNB-width at given contrast	spatial	Sobel	human-like	–	yes
LPC	auto-focus, image restoration, compression	weighted average of local phase coherence measures	wavelet	no	–	–	no
CPBD	multimedia content	percentage of edges at which blur can not be detected (in JNB sense)	spatial	Sobel	human-like	–	yes
FISH	general perceived sharpness	weighted average of log-energies of WT subbands	wavelet	no	–	–	no
S3	consumer photographs	slope of the magnitude spectrum combined with total spatial variation	spatial, Fourier	no	–	human-like	yes
SpaQF	auto-focusing, IQA, image fusion	combined local maxima of wavelet coefficients from different scales	wavelet	dedicated	–	minor	no
CogACR	general	average cone ratio of wavelet coefficients for the sharpest edges	wavelet	wavelet thresholding	yes	very low	no

measure applied directly to the images (no dictionary involved). These experiments focus on blur identification in multiply distorted images, corrupted simultaneously by blur (either GBlur, DBlur, or MBlur) and (Gaussian white) noise. We consider two levels of noise: $\sigma_{n1} = 10$ and $\sigma_{n2} = 25$.

Lastly, in Section 5.8.6, we make some remarks concerning the practical implementation of the CogACR methods on a commercial platform. Those results suggest that the proposed measure can achieve real-time performance for high-definition (HD) input.

5.8.1 Test image data

Test images in our experiments belong to three public databases: one non-IQ specific database of natural scene images prepared by [Oliva and Torralba, 2001], and two databases designed specifically for IQA which comprise both distortion-free and distorted images as well as corresponding human ratings of IQ – the LIVE Image Quality Assessment Database [Sheikh et al., 2006] and the very recent LIVE Multiply Distorted Image Quality Database [Jayaraman et al., 2012]. Details are described next.

5.8.1.1 Oliva & Torralba Environmental Scene Database (OlivaTorralba)

This established image collection has been proposed by [Oliva and Torralba, 2001] in the context of computational modelling of the recognition of real world scenes. The images come from the Corel stock photo library, pictures taken from a digital camera and images downloaded from the web. Specifically, the collection we use includes a total of 2688 images covering a variety of outdoor places, natural scenes[13] as well as urban environments. The images are grouped in 8 contextual categories (databases): coast and beach (DB1), forest (DB2), highway (DB3), city center (DB4), mountain (DB5), open country (DB6), street (DB7), and tall building (DB8); example images are depicted in Figure 5.8.3. Each category consists of several hundreds of color images of 256×256 pixels in size, in 256 gray levels. The details about size of each database (N_{REF}) are mentioned in Table 5.2. Hereafter, we refer to this database as "OlivaTorralba".

We use these images to study the CogACR performance for NR defocus estimation in Section 5.8.4. For our specific application, the database has two main advantages: it includes a large number of images to accommodate the requirements of our training and testing processes and it groups the images based on their content. The latter aspect has a two-fold benefit: firstly, it ensures reasonably representative training data (as long as we stay within a given image category), and secondly, it allows looking at the influence of image content on the performance of our measure. Details are discussed later in Section 5.8.4.

[13] The database was used by [Moorthy and Bovik, 2010] to assess statistics of natural image distortions.

5.8.1.2 LIVE Image Quality Assessment Database Release 2 (LIVE1Blur)

The LIVE database is definitely among the most referenced ones in the literature of IQA. It has been published by [Sheikh et al., 2006]. The database was derived from a set of 29 undistorted (REF) high resolution and high quality color images collected from the web and photographic CD-ROMs. Though not exhaustive, the images include a variety of content, as shown in Figure 5.22. Most images are of the order of 768×512 pixels in size. For our experiments, we use the images distorted by GBlur created by filtering each R, G, and B components of the REF images by a circular-symmetric 2D Gaussian kernel of standard deviation σ pixels. The three color components of the image were blurred using the same kernel determined by $\sigma \in [0.42, 15]$ pixels. Hereafter, we refer to this database as "LIVE1Blur".

The quality ratings of humans were collected in single-stimulus experiments (SS) [ITU-R, 2012] where each image was rated by an average number of 24 subjects (digital image/signal processing students). The reported subjective judgments of quality were linearly mapped to the interval [1,100]. The images were displayed on CRT monitors at a resolution of 1024×768 pixels. The results of human experiments are released in the form of the difference mean opinion scores (DMOS)[14] and their standard deviation; raw scores assigned to the images by individual human observers are not provided.

Images from the LIVE database are used throughout this chapter, either to illustrate the properties of the discussed techniques (in the previous sections) or to test the performance of the proposed methods (in Section 5.8.3) and where applicable to compare them to humans (in Section 5.8.3). Note that human ratings of the IQ were collected for color displayed images. In all our experiments, the images are used in grayscale representation.

5.8.1.3 LIVE *Multiply Distorted* Image Quality Database (LIVE2BlurNoise)

This is the most recent public database created for the purpose of benchmarking objective IQ assessment algorithms. Unlike the previous databases which considered different types of distortions always one at a time, the new database is comprised of images with *simultaneous multiple distortions*. In particular, we consider Part2 of the database which contains images corrupted jointly by GBlur noiseive GWN [Jayaraman et al., 2012]. For each of the two distortions, three different parameter values were considered: $\sigma_b \in \{3.2, 3.9, 4.6\}$ and $\sigma_n \in \{11.5, 22.8, 45.6\}$. Thus, the database contains a total of 240 images: 15 REFs and 225 distorted images of which 90 with a single distortion (45 of each GBlur and GWN) and 135 with multi-distortion (15 for each

[14] The mean opinion score (MOS) is defined as the mean of observer's individual opinion scores, *i.e.*, of the values on a predefined scale that observers assign to their opinion of the overall quality of the image. The difference mean opinion score (DMOS) is the difference between the MOS of the test image and that of the corresponding undistorted (REF) image, DMOS = MOS(REF image) - MOS(test image).

Figure 5.22: The 29 REF images from the LIVE1Blur database.

of the 9 possible combinations of the GBlur and GWN parameters). The images are 1280×720 pixels in size. Hereafter, we refer to this database as "LIVE2BlurNoise".

Together with the images, the database Part2 provides the perceptual IQ scores of a total of 18 human observers, mostly volunteer graduate students, collected using a single stimulus with hidden reference removal method[15] [Pinson and Wolf, 2003] with a continuous scale from 0 to 100. The images were displayed on an LCD monitor at a resolution of 1280×720 pixels. The results of human experiments are released in the form of raw scores assigned to the images by individual human observers. Figure 5.23 depicts the 15 REFs and their corresponding MOS scores.

We use this database in Section 5.8.5 to test the performance of 12 state-of-the-art NR blur measures (including the CogACR proposed in this chapter) for images which are not only blurry but also corrupted by noise (which is typically the case in practice). Note, here also, that human scores are collected for color displayed images while our experiments are with grayscale data.

5.8.2 Edge detection

In order to evaluate the performance of the technique for edge detection proposed in Section 5.3.2, we apply it to five different image contents (see the images in row 1 of Figure 5.24): "Butterfly", "FishingBoat", "Houses", "Man", and "Peppers". The latter four images were used also in [Ferzli and Karam, 2009] for evaluation of the state-of-the-art NR image blurriness measures. We note from Figure 5.24 that the five REF images are quite diverse in content: the "Houses" image contains many sharp edges, the same as the "FishingBoat" image only here also smooth areas are present (water, sky), the "Man" image is quite rich in texture (feathers, clothes) in contrast to the "Peppers" image which contains mostly smooth regions. Finally, we include the "Butterfly" image as a representative of the content with varying level of blurriness between the object and the surrounding areas. For the purpose of experiments, the REF images are distorted (1) by DBlur filtering parametrized by $r = \{3, 7\}$ pixels only, and (2) by DBlur filtering followed by adding GWN of $\sigma_{n1} = 10$ and $\sigma_{n2} = 20$. The distorted images are also depicted in Figure 5.24.

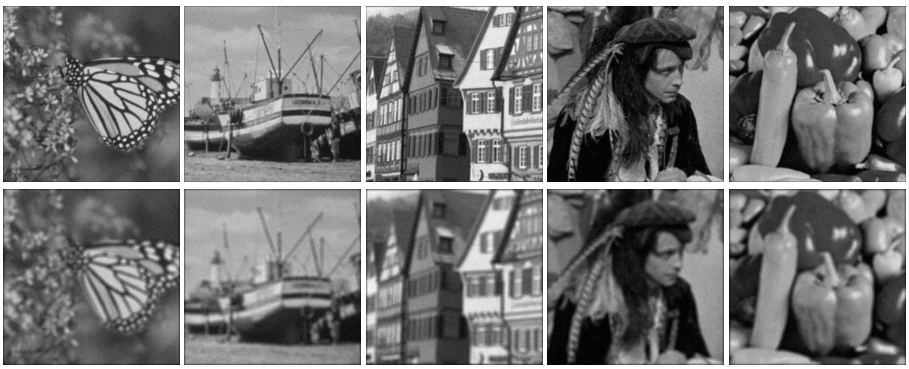
Figure 5.25 and Figure 5.26 show the results of, respectively, the Sobel edge detection (often used in the blur estimation techniques, see Table 5.1 for some example methods) and the proposed method described in Section 5.3.2. It is easy to note from the Sobel edge maps that the method is sensitive to image blur and even more so to image noise, neither of which is a desired property of an edge detection method. To its advantage, the proposed method appears quite immune to both blur and noise in the image.

[15] According to [Pinson and Wolf, 2003], the hidden reference removal refers to the strategy in which the reference video sequences are presented during the test session, but observers are not aware that they are evaluating the reference image.

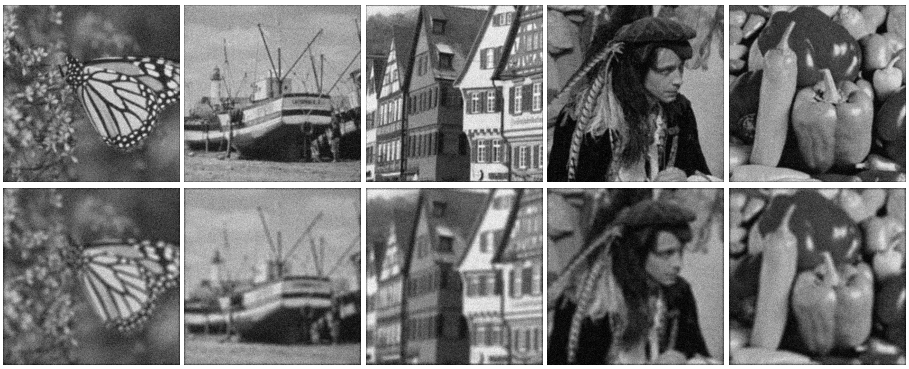
Figure 5.23: The 15 REF images from the LIVE2BlurNoise database are shown together with their names and their mean opinion scores (MOS). The images are arranged according to their MOS values, from the highest (top left) to the lowest score (bottom right).

REF images (from left to right): Butterfly, FishingBoat, Houses, Man, Peppers

Defocus blurred images: (top) r = 3, (bottom) r = 7

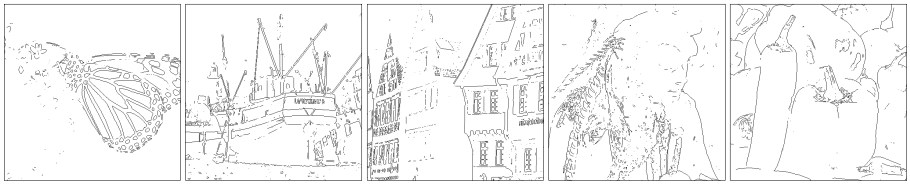


Defocus blurred images with added GWN: σ_{n1}=10

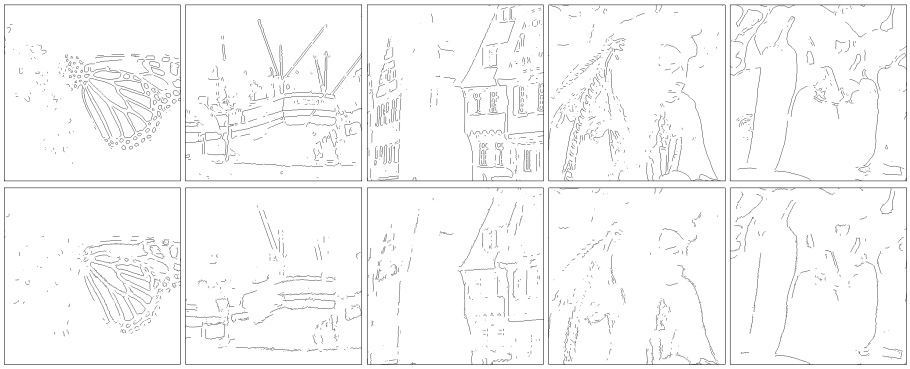


Defocus blurred images with added GWN: σ_{n2}=25

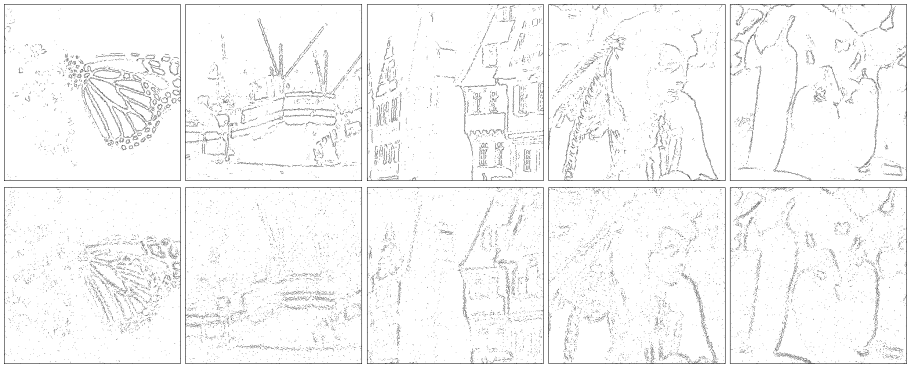
Figure 5.24: The five test images and their DBlur and GWN distorted versions.



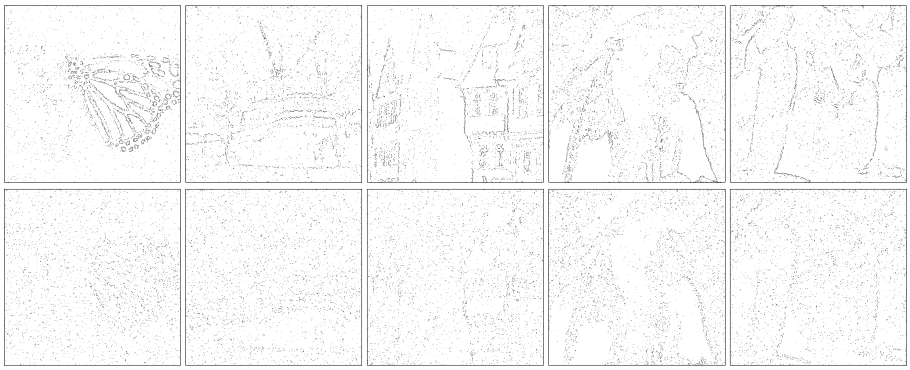
REF images (from left to right): Butterfly, FishingBoat, Houses, Man, Peppers



Defocus blurred images: (top) r = 3, (bottom) r = 7



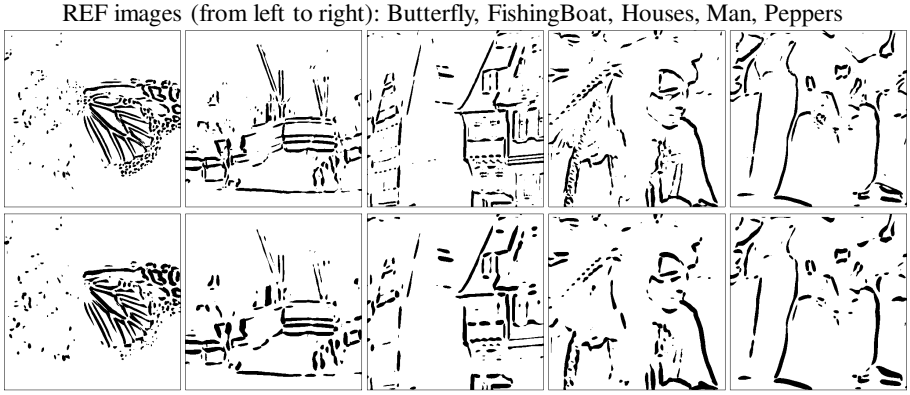
Defocus blurred images with added GWN: σ_{n1}=10



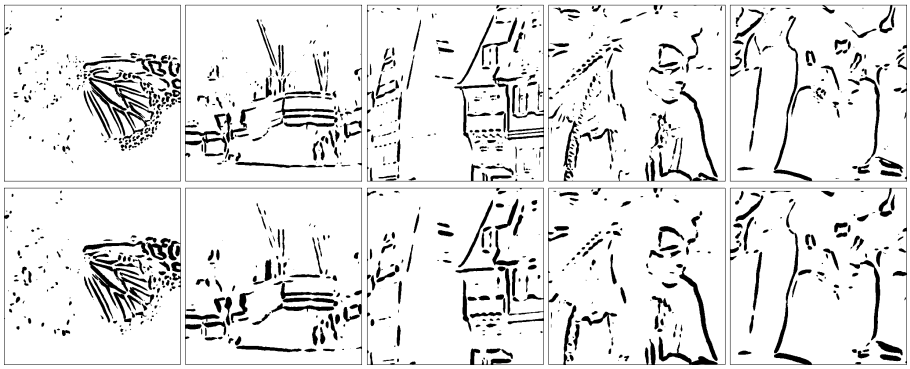
Defocus blurred images with added GWN: σ_{n2}=25

Figure 5.25: For the images from Figure 5.24, edge maps detected by Sobel detector.

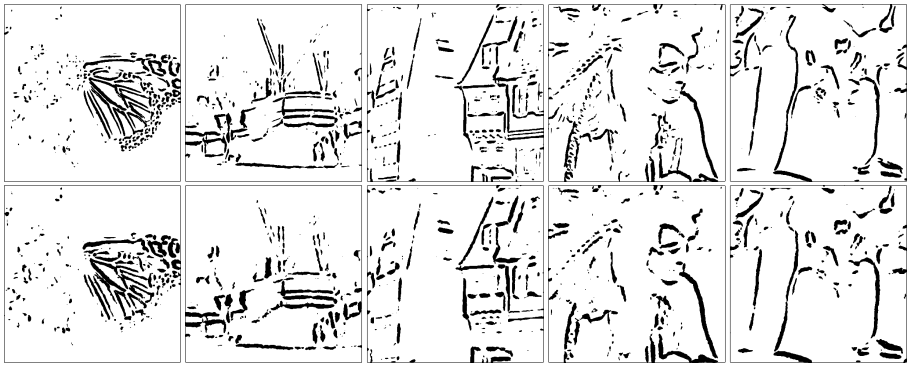
REF images (from left to right): Butterfly, FishingBoat, Houses, Man, Peppers



Defocus blurred images: (top) r = 3, (bottom) r = 7



Defocus blurred images with added GWN: σ_{n1}=10



Defocus blurred images with added GWN: σ_{n2}=25

Figure 5.26: For the images from Figure 5.24, edge maps detected by the method proposed in Section 5.3.2. We use the wavelet inter-scale product $\mathrm{P}_{3\to4}$ and the threshold percent index $\mu = 5\%$.

(a) d_{SKL}= 0.0126 (b) d_{SKL}= 2.9124

Figure 5.27: Undistorted images from LIVE1Blur database were compared for their ACR histograms, as explained in Section 5.5. (a) Two most similar images and (b) two most dissimilar images are shown together with the corresponding values of the dissimilarity index d_{SKL} (larger value of d_{SKL} corresponds to larger dissimilarity, *i.e.*, lower similarity).

5.8.3 Best matching images

We conduct several experiments to evaluate the performance of the HistACR edge descriptor proposed in Section 5.5. First, according to the proposed edge-based method of assessing image similarity, we find the two most and the two least similar images among the 29 REFs from LIVE1Blur database. All 29 REF images are shown in Figure 5.22 and the selected two pairs of the most and the least similar among those are depicted in Figure 5.27. The images of "Statue" and "CarnivalDolls" are suggested by the method to have the most similar edge content. Indeed, the two images appear similar for their edges also visually – they both have many sharp not very long edges. Likewise, the two images suggested least similar "StudentSculpture" and "Parrots" also visually seem rather different; for one thing, the background of "Parrots" is rather smooth (blurry) while the background of "StudentSculpture" contains many small edges (texture-like).

Next, in Figure 5.28, we compare the CogACR behaviour of the suggested image pairs for different levels of GBlur. Remember that the experiments are done for the images from LIVE1Blur database, thus only the corresponding levels of GBlur are considered. As expected, the CogACR values of the two most similar images agree

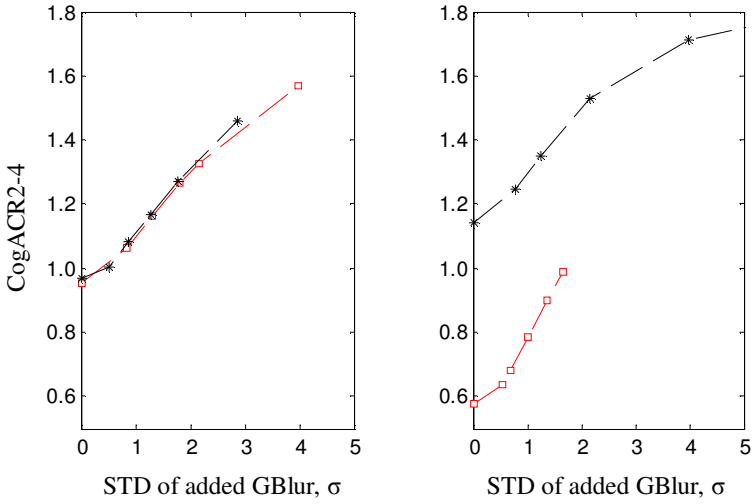


Figure 5.28: CogACR24 values computed for the two pairs of REF images in Figure 5.27 and their GBlur distorted variants from LIVE1Blur image database. Note that the CogACR24 values corresponding to Figure 5.27 (a) are very similar while those corresponding to Figure 5.27 (b) are obviously different.

very well (left plot in Figure 5.28) unlike those of the two least similar images (right plot in Figure 5.28).

Lastly, we examine also the agreement between human ratings of IQ for two image pairs. Perhaps less expectedly, also the human MOS values seem to agree with the proposed method, *i.e.*, humans rated very similarly the quality of blurred "Statue" and "CarnivalDolls" images (left plot in Figure 5.29), and they rated rather differently the quality of blurred "StudentSculpture" and "Parrots" images (right plot in Figure 5.29).

5.8.4 NR defocus estimation

Originating from focus problems with digital cameras and lens aberrations, the DBlur frequently appears in the digital images of natural scenes, especially when these are taken with non-specialized cameras and by non-professional photographers. Moreover, estimating the amount of defocus is also of interest for depth estimation in the image [Levin, 2007].

In order to objectively evaluate the performance of the CogACR measure in the task of estimating the radius of DBlur, we conduct a range of experiments using the images of the OlivaTorralba database described in Section 5.8.1.3. The original color images are converted to grayscale; these are considered the REFs (degradation-free data). To create blur-distorted images, we introduce DBlur to the REFs using the DBlur model from Eq. ((5.3)). Each image is distorted using $N_{\mathrm{BL}} = 40$ different values of blur radius varied in the range from 0.25 to 10.00 pixels and with a uniform

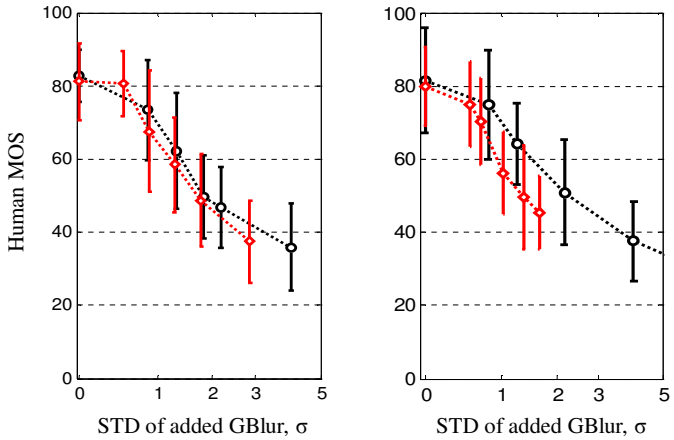


Figure 5.29: Human MOS values corresponding to the two pairs of REF images in Figure 5.27 and their GBlur distorted variants from LIVE1Blur database. Error bars are ± 2 standard deviations of the MOS values. The MOS values and their standard deviation are taken from the "LIVE Gaussian Blur Test Image Set" downloaded from `https://ivulab.asu.edu/software/quality/cpbd`. Note that the human MOS corresponding to Figure 5.27 (a) are very similar while those corresponding to Figure 5.27 (b) are clearly different.

step of 0.25, thus $r_m = 0.25\,m, m = 1, \ldots, N_{\mathrm{BL}}$.

In these experiments, the CogACR measure is computed following the procedure described in Section 5.6 and depicted in Figure 5.21. Thus, for a given test image, we estimate the level of blur BL by searching for its best match among the training images. Then, we take the known BL of the best matching training image as an estimate of BL of the test image.

The experiments are conducted as follows. First, we split each of the 8 categories (DBs) of the OlivaTorralba images, in two a randomly selected subsets: the *training* subset of $N_{\mathrm{tr}} = 200$ REF images and the *testing* subset of $N_{\mathrm{ts}} = 50$ REFs. There is no overlap between the training and the testing subsets. Then, we use the training images to build the dictionary. Once the dictionary is available, we take a test image and employ the method from Section 5.5 to select the candidates and find the best matching element from the dictionary. Finally, the known level of blur of the best matching dictionary element is taken as the estimated level of blur for the test image at hand. This search process is repeated for each image from the testing subset.

For illustration of the performance of the best matching method, Figure 5.30 presents the best matches found among the REF images. Note, however, that in the actual experiments the REF images are not considered separately but they are treated in the same way as the distorted images. In particular, the pairs of images shown in Figure 5.30 are the most similar images found separately for each of the 8 DBs.

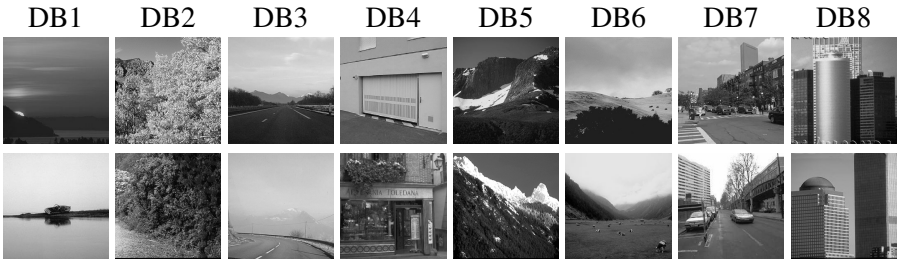


Figure 5.30: The best match image pairs for each of the 8 image databases used in the study. The images are taken from the image collection proposed in [Oliva and Torralba, 2001] which consists of 2688 color natural scene images split in 8 categories: coast and beach (DB1), forest (DB2), highway (DB3), city center (DB4), mountain (DB5), open country (DB6), street (DB7), and tall building (DB8). All images are converted to gray scale and kept in the original size of 256×256 pixels.

Among the shown pairs, the similarity between the best matching images is observed largest within DB2 (Forest) and DB5 (Mountain), and smallest within DB1 (Coast & Beach), DB6 (Open Country) and DB8 (Tall Building).

Now, we evaluate the accuracy of the estimated blur levels. For the purpose of comparison, we consider a range of state-of-the-art NR blur measures described in Section 5.7. In total, the experiments include twelve NR blur measures: the ten existing measures – CPBD, EdgeWidth, FISH, JNBM, Kurt2-Freq, Kurt2-WV, LPC, EdgeAPR, S3, and SpaQF, and the two variants of the proposed measure – CogACR24 and CogACR34. In addition, we consider the PSNR measure which is still the most commonly used IQ in the field of image processing; note though that PSNR is a full-reference measure. The experiments are carried out for each DB1 to DB8. The REF images are pooled together with their blur distorted realizations, thus there was a total of $N_{\mathrm{REF}}(N_{\mathrm{BL}}+1)$ images per database.

The measures are evaluated in terms of the Spearman rank-order correlation coefficient (SROCC) computed between the measure values and the known ("true") blur radii of an image. The SROCC statistic indicates monotonicity of the measure and is commonly used for evaluation of the IQ methods. High values of SROCC (close to 1) indicate high accuracy of the predictions made by the method.

The SROCC values for all tested methods are summarized in Table 5.2. Based on these results, the CogACR24 ranks among the three best performing measures, together with FISH and LPC. This hold for all 8 categories of image content.

Furthermore, we perform an extensive experimental study to evaluate the effects of parameters of the proposed CogACR based method for NR blur identification. The study involves a range of experiments performed for each DB separately. Each experiment is characterized by the following list of parameters: training set $\mathcal{X}$ of the size N_{tr}, testing set $\mathcal{Y}$ of the size N_{ts}, the number of the considered blur levels N_{BL} (different blur radii), and the value of the threshold parameter T_{C}.

Table 5.2: Comparison of the performance of 12 NR blur distortion measures (see Section 5.7 for details). In addition, the correlation coefficients for PSNR, the most frequently used FR measure, are also shown. The evaluation is done using the SROCC. The calculations are performed for each of the 8 image categories using a total of $N_{\mathrm{REF}}(N_{\mathrm{BL}}+1)$ images per data base: N_{REF} blur-free images together with $N_{\mathrm{BL}}=40$ of their DBlur distorted copies.

	DB1	DB2	DB3	DB4	DB5	DB6	DB7	DB8
N_{REF}	360	328	260	308	374	410	292	356
CogACR24	0.8881	0.9518	0.9407	0.9535	0.9412	0.9364	0.9619	0.9509
CogACR34	0.8298	0.9084	0.8948	0.9041	0.9000	0.8900	0.9237	0.9081
CPBD	0.8385	0.9173	0.8762	0.8885	0.9172	0.8799	0.9335	0.8634
EdgeWidth	0.3715	0.3194	0.2929	0.0733	0.2355	0.1920	0.2864	0.3327
FISH	0.8984	0.9528	0.9544	0.9585	0.9358	0.9285	0.9764	0.9384
JNBM	0.5772	0.9126	0.7189	0.9052	0.8712	0.7446	0.9301	0.9154
Kurt2-Freq	0.1572	0.2816	0.2554	0.4170	0.2218	0.2597	0.2583	0.2882
Kurt2-WV	0.2104	0.1200	0.2453	0.2969	0.2631	0.2289	0.2520	0.3307
LPC	0.8926	0.9458	0.9479	0.9419	0.9373	0.9337	0.9678	0.9242
EdgeAPR	0.8259	0.7371	0.8104	0.8606	0.8773	0.8211	0.8898	0.8484
S3	0.8160	0.8891	0.8825	0.9089	0.8610	0.8652	0.9334	0.8805
SpaQF	0.7889	0.9289	0.8749	0.8976	0.9196	0.8840	0.9364	0.9176
PSNR	0.2359	0.1919	0.3077	0.3581	0.2537	0.1967	0.3790	0.2774

First, from each DB1 to DB8, we select two random subsets of images to build the training and the testing set, $\mathcal{X}$ and $\mathcal{Y}$, respectively. These data splits are done such that there is no overlap between the two subsets. Also, note here that in these experiments the REF and its corresponding blur distorted images are kept together. Hence, the total number of images in a subset is $N_{\mathrm{REF}}(N_{\mathrm{BL}}+1)$. This considered, the size of each training set is $N_{\mathrm{tr}}=N_{\mathrm{REFtr}}(N_{\mathrm{BL}}+1)$ and the size of each testing set is $N_{\mathrm{ts}}=N_{\mathrm{REFts}}(N_{\mathrm{BL}}+1)$. In particular, we consider the following parameter values: $N_{\mathrm{REFtr}}\in\{50,100,150,200,250,300,350\}$ where the maximum value for a the given DB depends on its size, $N_{\mathrm{REFts}}=50$, while the number of BLs is fixed at $N_{\mathrm{BL}}=40$.

Next, for each training set $\mathcal{X}$, we compute the ACR histograms and build the dictionary $\mathcal{D}$. This completes the set of necessary inputs for the trained NR blur estimation algorithm described in Section 5.6 and Figure 5.21.

Now, we run the NR blur estimation procedure. For each of the N_{ts} test images, we extract the candidate set of entries and find the best matching candidate. The radius of blur corresponding to the best match candidate is the estimated radius of blur for the given test image, r. In these experiments, the value of the threshold T_{C} used in selecting the subset of candidates is varied among three values, $T_{\mathrm{C}}\in\{0.00005,0.0005,0.005\}$. All three T_{C} values are considered for each dictionary $\mathcal{D}$ and each testing image set $\mathcal{Y}$.

The results of all experiments for the largest among 8 image DBs, DB6 (Open Country), are depicted in Figure 5.31. In the performance analysis of the proposed NR blur measure, we use as a figure of merit the absolute error between the estimated blur radius and the actual blur radius, $\Delta = |r - r^*|$. Here, r is the estimated value of blur radius and r^* is the actual value of blur radius. There, we distinguish three conditions: (1) $\Delta \leq 1$ pixel, (2) 1 pixel $< \Delta \leq 2$ pixels, and (3) $\Delta > 2$ pixels. The top, middle and bottom plots in Figure 5.31, respectively, correspond to these three ranges of the absolute error in blur radius estimation. Different lines in each plot represent different value of the threshold T_{C}.

Importantly, in order to avoid bias of the results and allow for their statistical significance to be evaluated, for each parameter setup of N_{REFtr} and N_{REFts}, we explore 5 different random realizations of data grouping, training versus testing data. We then use the scores from these 5 realizations to estimate the error bars for our results. In Figure 5.31, these are shown as $\pm$ one standard deviation of the scores across 5 random initializations of each experiment.

The results for DB6 indicate that the highest and approximately stable performance of the NR blur measure is achieved for $N_{\mathrm{REFtr}} = 200$ and $T_{\mathrm{C}} = 0.005$. In Figure 5.32, we use these parameters to illustrate the performance of the method across 8 different image DBs. The results are shown as a bar chart where colors represent different ranges in accuracy of the method, similar as in Figure 5.31, and the error bars extend to $\pm$ one standard deviation of the scores across 5 random initializations of each experiment. Accordingly, the results from Figure 5.32 indicate that in 6 out of the 8 DBs the absolute error between the estimated and the true blur radius is $\Delta \leq 1$ pixel in 85% or more of the considered cases of blur estimation. In addition, for all 8 DBs, the percentage of the cases in which $\Delta > 2$ pixels remains below 5%, most often below 3%.

Related to the proposed method for NR blur identification and the threshold parameter of the candidate selection, T_{C}, we noted from the results shown in Figure 5.31 that the greater T_{C} allows higher performance of the method. We remind that the greater T_{C} means less strict condition of the CogACR based similarity, d_{E}, as defined in Eq. 5.20. Then, the observed trend of greater T_{C} resulting in greater accuracy of the blur estimation, may suggest that, once we are in the right neighborhood of CogACR values, T_{C} neighborhood, the finer details of the ACR histogram become of high importance. That is exactly the point in the algorithm where we measure the similarity of ACR histograms, d_{SKL}, as defined in Eq. 5.18. Again, this increased significance of all components of the ACR histogram, rather than only its center of gravity, point to the strong relationship between the content of the image and the effect of blur that can be measured or even perceived.

To end with, one more interesting aspect to discuss is the influence of size and structure of the dictionary on the performance of the proposed NR blur estimation. While always relatively high, we notice in Figure 5.32 that the performance of the measure slightly varies among different image categories. One reason for this could

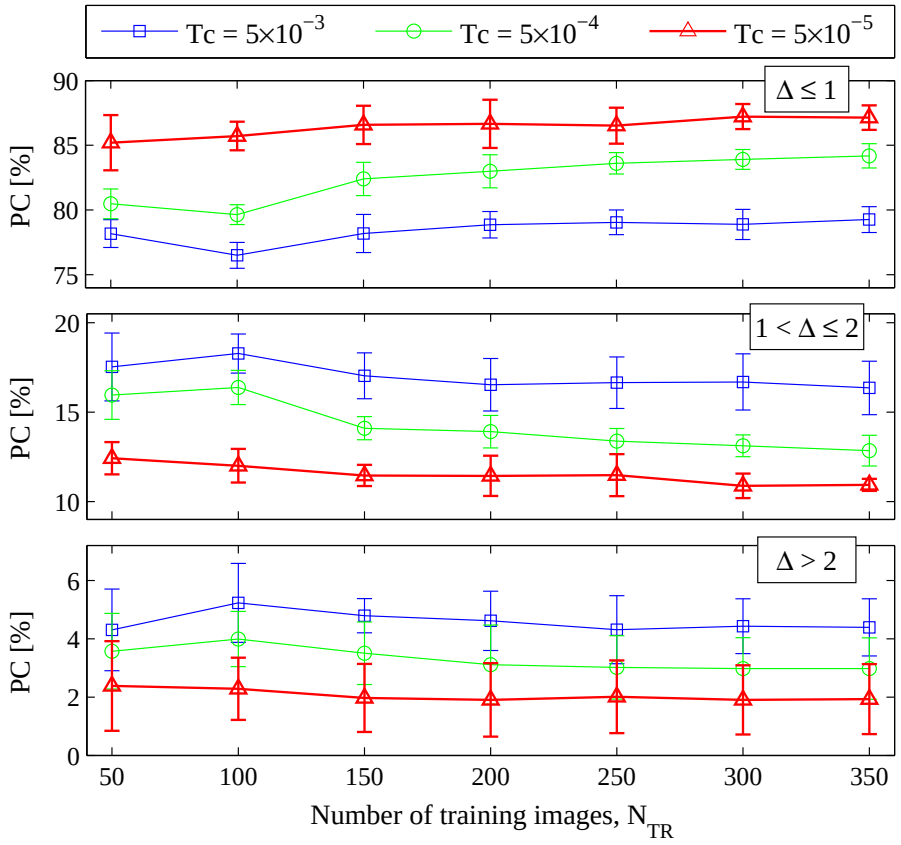


Figure 5.31: Performance of the proposed NR blur estimation algorithm evaluated by comparing the percent correct (PC) on 7 different sizes of the training dataset, with N_{REFtr}={50, 100, 150, 200, 250, 300, 350} and on 3 different threshold values for candidate selection, T_C={0.00005, 0.0005, 0.005}. The number of reference test images is kept constant in all experiments, $N_{\mathrm{REFts}} = 50$, while the number of different blur radii is $N_{\mathrm{BL}} = 40$. The three plots differ in the value of absolute error Δ of the estimated blur radius r: (top) $\Delta \leq 1$ pixel, (middle) $1\mathrm{pixel} < \Delta \leq 2$ pixels, (bottom) $\Delta > 2$ pixels. The scores are shown for DBof Open Country (DB6). Error bars indicate the $\pm$ standard deviation of the scores across 5 different random selections of training and testing image subsets. In each of the 5 selections, no overlap exists between the trainer images and the tester images.

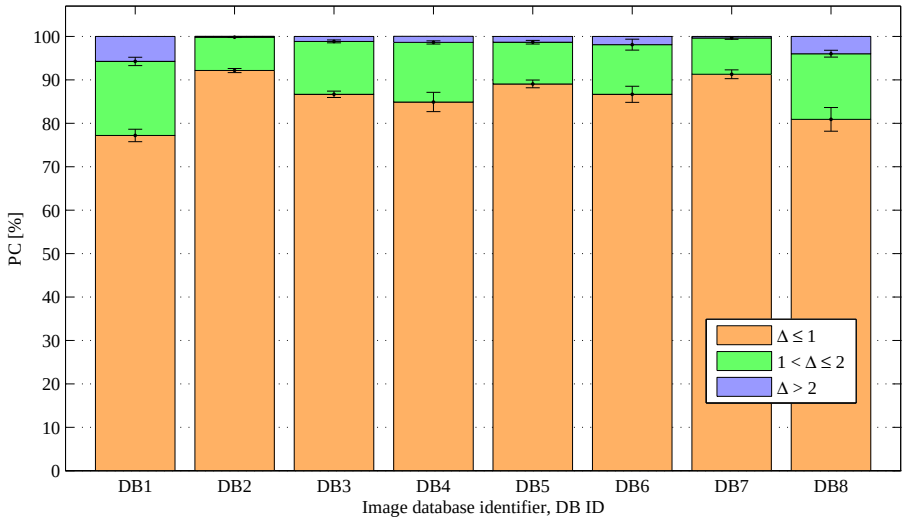


Figure 5.32: Performance of the proposed NR blur estimation algorithm evaluated by comparing the percent correct (PC) on all eight DBs. For each DB, the following parameter values are used: $N_{\mathrm{REFtr}} = 200$, $N_{\mathrm{REFts}} = 50$, $N_{\mathrm{BL}} = 40$, and T_{C}=0.005. For each data category, 5 different random selections of training and testing image subsets are considered. In each experiment, no overlap exists between the trainer images and the tester images. Error bars are the $\pm$ standard deviation of the scores across 5 data groupings.

be different sensitivity of the measure to different types of image content, as discussed earlier. In that sense, we note that the NR blur measure approximately follows the trends in correlation coefficient of CogACR: the higher CogACR correlation, the higher percent of cases with high estimation accuracy, $\Delta > 1$.

Nevertheless, the oscillation in algorithm performance could also be due to the effect of the dictionary. We noted in Figure 5.31 that the size of dictionary affects the performance of the method, the larger the dictionary the higher the performance, especially in the range of "smaller" dictionaries ($N_{\mathrm{REFtr}} \leq 200$, that is $N_{\mathrm{tr}} \leq 8200$). Next to its size, it is very important to know how well and how complete the dictionary represents the data. In case of our method, one way to anticipate this would be looking at the statistics of d_{SKL} obtained for the testing images. For example, in the experiments from Figure 5.32, the mean value of the d_{SKL} across all images in the testing set of the given DB (not shown here) are exactly proportional to the accuracy of the method in that DB: the higher the accuracy of the method, the smaller the average distance between the test image and its best match.

5.8.5 CogACR performance in noise corrupted images

In these experiments, we evaluate the performance of the proposed CogACR measure of image blurriness in the condition where images are corrupted also by large amount of noise. The same as previously, all experiments involve comparisons with the best-performing existing NR blur measures from the literature (see Section 5.7). Firstly, we examine the performance of the selected blur measures for five example REF images (each of a different kind of content). The images are distorted by three types of image blur (GBlur, DBlur, and MBlur) and two levels of image noise ($\sigma_{n1} = 10$ and $\sigma_{n2} = 25$). Secondly, we apply the measures on an existing public database of multiply distorted images LIVE2BlurNoise (GBlur and Gaussian white noise) and evaluate the correlation between the measure scores and (a) the true amount of blur, as well as (b) the MOS values of humans. The details are presented next.

5.8.5.1 Evaluation for different types of blur

The results reported in this section are performed for the five REF images used also in Section 5.8.2: : "Butterfly", "FishingBoat", "Houses", "Man", and "Peppers". The REF images are distorted by three types of blur: GBlur, DBlur, and MBlur, always one at a time. The corresponding blur models are defined by Eq. ((5.3)), Eq. ((5.3)), and Eq. ((5.3)), respectively. The same as in the previous experiments, we use the blur models to create blurred variants of the REF images. The level of introduced blurring covers the range described by $\mathrm{BL} = 0.25\,m, m = 1, \ldots, N_{\mathrm{BL}}$ where BL refers to the relevant blurring filter parameter: the standard deviation σ of the GBlur, the radius r of the DBlur, and the length l of the MBlur. Moreover, the REF as well as the blurry images are distorted by adding GWN of $\sigma_{n1} = 10$ and $\sigma_{n2} = 25$. The experiments also include $\sigma_n = 15$; however, the details are omitted here as those results are not especially indicative of the tested behaviour.

The lower level of noise: $\sigma_{n1} = 10$

In order to describe the degree of the considered image distortion on a more "conventional" scale, we show the PSNR values in Figure 5.33. Overall, the PSNR values cover the range from about 40 dB (images with the lowest level of blur) down to about 15 dB (images with the highest level of blur). Note also that PSNR values of the blur-free but noisy images start at below 30 dB.

We assess the proposed measures CogACR24 and CogACR34 with regards to their agreement with the true parameters of blur, both in the absence of noise and for the noisy images. The same as in the previous Section 5.8.4, the proposed methods are evaluated in comparison to the existing NR methods of CPBD, EdgeWidth, FISH, JNBM, Kurt2-Freq, Kurt2-WV, LPC, EdgeAPR, S3, and SpaQF. The results are depicted in Figure 5.34.

In the case where there is no added noise in the images (colored markers in the plots), all but a few measures exhibit monotonicity for GBlur and DBlur. The ex-

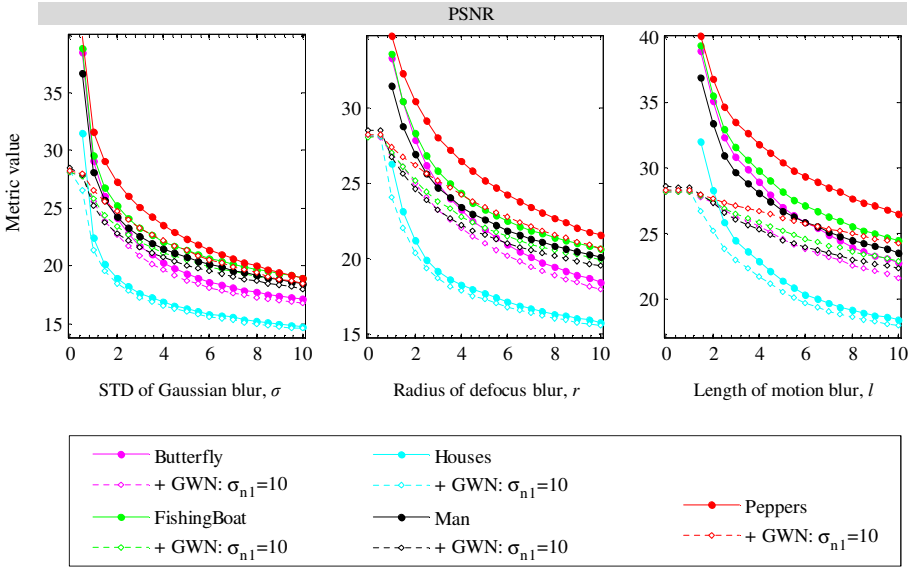


Figure 5.33: Performance of the full-reference PSNR measure for images corrupted with noise of $\sigma_{n1} = 10$. The plots correspond to three different types of image blur (from left to right): GBlur, DBlur, and MBlur. Different colors represent different image content ("Butterfly", "FishingBoat", "Houses", "Man", and "Peppers"), solid lines correspond to the noise-free images, and dashed lines to the images with noise of $\sigma_{n1} = 10$

ceptions are EdgeWidth, which behaves nonmonotonic for all three types of blur, and LPC and EdgeAPR, which lose monotonicity only for DBlur. Nevertheless, only a few measures preserve monotonicity properties also for MBlur distortion (Kurt2-Freq, Kurt2-WV, CogACR24, CogACR34). On the other hand, in the presence of noise, the majority of the methods drastically changes their behaviour, irrespective of the image content and of the type of blur (compare the white markers in the plots versus the colored ones). The methods which exhibit little sensitivity to the noise of $\sigma_{n1} = 10$ are the same four measures that exhibit monotonicity for all three types of blur: Kurt2-Freq, Kurt2-WV, CogACR24, CogACR34. Among these four measures, Kurt2-Freq appears the least immune to noise.

The higher level of noise: $\sigma_{n2} = 25$

Then, we evaluate the selected four measures for higher levels of noise. No significant changes are observed for the noise of $\sigma_n = 15$ (details not included here). The next higher level of noise is $\sigma_{n2} = 25$; these results are shown in Figure 5.35. Overall, based on the plots, the proposed CogACR34 measure remains nearly unaffected by noise even at this high level (except for the high levels of GBlur). The second best is the Kurt2-WV measure (except for the high levels of GBlur where it performs best).

Moreover, we examine the performance of the proposed CogACR24 and CogACR34 at $\sigma_{n1} = 10$ relative to that at $\sigma_{n2} = 25$. If we compare the two at $\sigma_{n1} = 10$ (see Figure 5.34), both measures are nearly unaffected by noise. However, at $\sigma_{n2} = 25$, CogACR24 is less immune to noise than CogACR34 (especially at higher levels of GBlur and DBlur). The reason for the observed inferior performance of CogACR24 is probably related to the accuracy of the corresponding edge maps (localization of edges) which gets affected by very high levels of noise. Remember that in CogACR24 we use wavelet scale 2^2, 2^3 and 2^4 to determine the edges while in CogACR34 we use only scales 2^3 and 2^4. This suggests that, at the very high level of noise, it may be better to not include coefficients of the scale 2^2 in the process of edge detection as they may still contain significant traces of noise. Instead, larger wavelet scales ought to be considered, *e.g.*, scale 2^3 and scale 2^4.

Lastly, related to the discussion about the effects of edge detection, we perform an experiment in which the Sobel edge maps of EdgeWidth method are replaced by the edge maps of CogACR34. These results are shown in the plots in Figure 5.37. By comparing these to the corresponding plots from Figure 5.34 obtained for the Sobel edge maps, we notice the great improvement in the sense of measure monotonicity (for noise-free images). This is another evidence of the importance of accurate edge detection in the process of blur identification.

5.8.5.2 Evaluation for multiply distorted images

In these experiments, we evaluate the performance of the proposed blur measures for a very recent public database LIVE2BlurNoise of multiply distorted images (GBlur

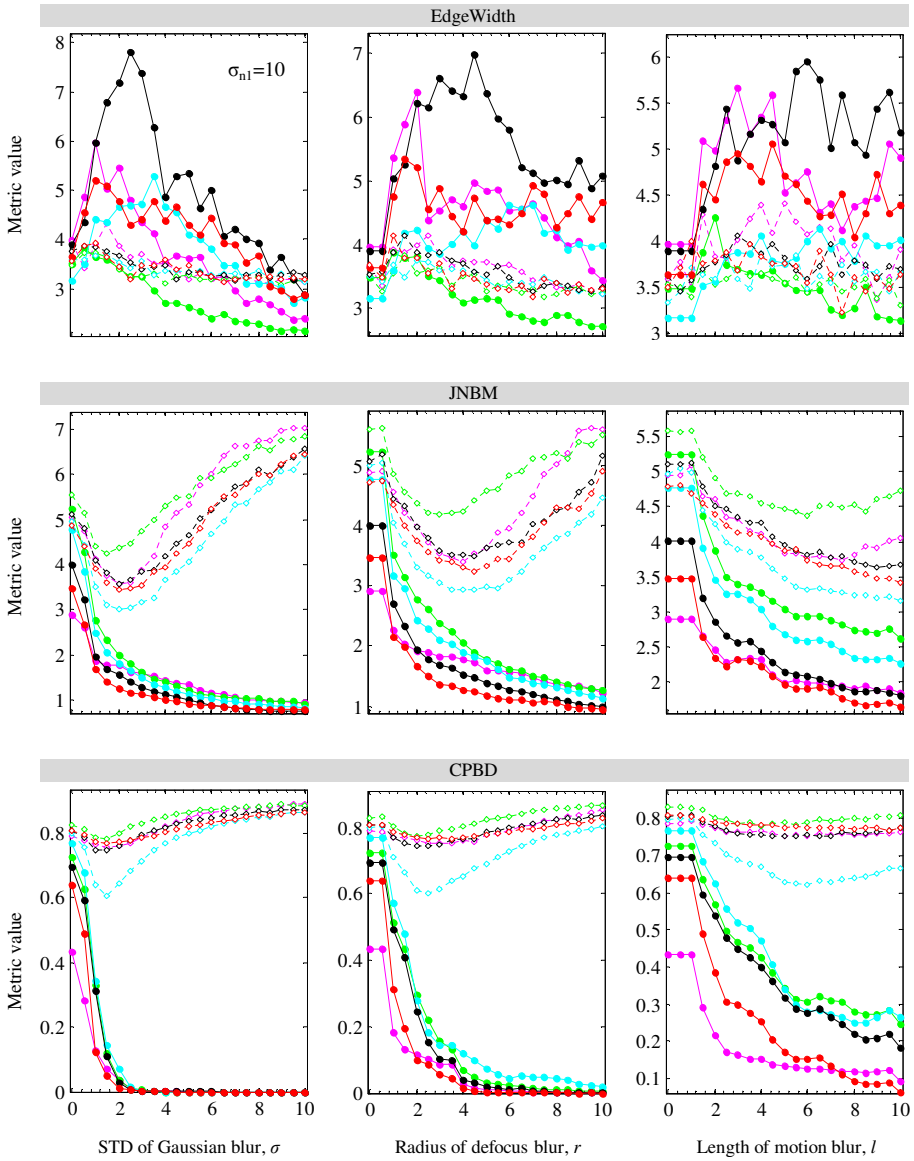


Figure 5.34: Evaluation of the NR blur measures for images corrupted with noise of $\sigma_{n1} = 10$. The columns correspond to three different types of image blur (from left to right): GBlur, DBlur, and MBlur. X-axis labels on the bottom apply across the entire column. The legend is the same as in Figure 5.33: different colors represent different image content ("Butterfly", "FishingBoat", "Houses", "Man", and "Peppers"), solid lines correspond to the noise-free images, and dashed lines to the images with noise of $\sigma_{n1} = 10$. (Continued)

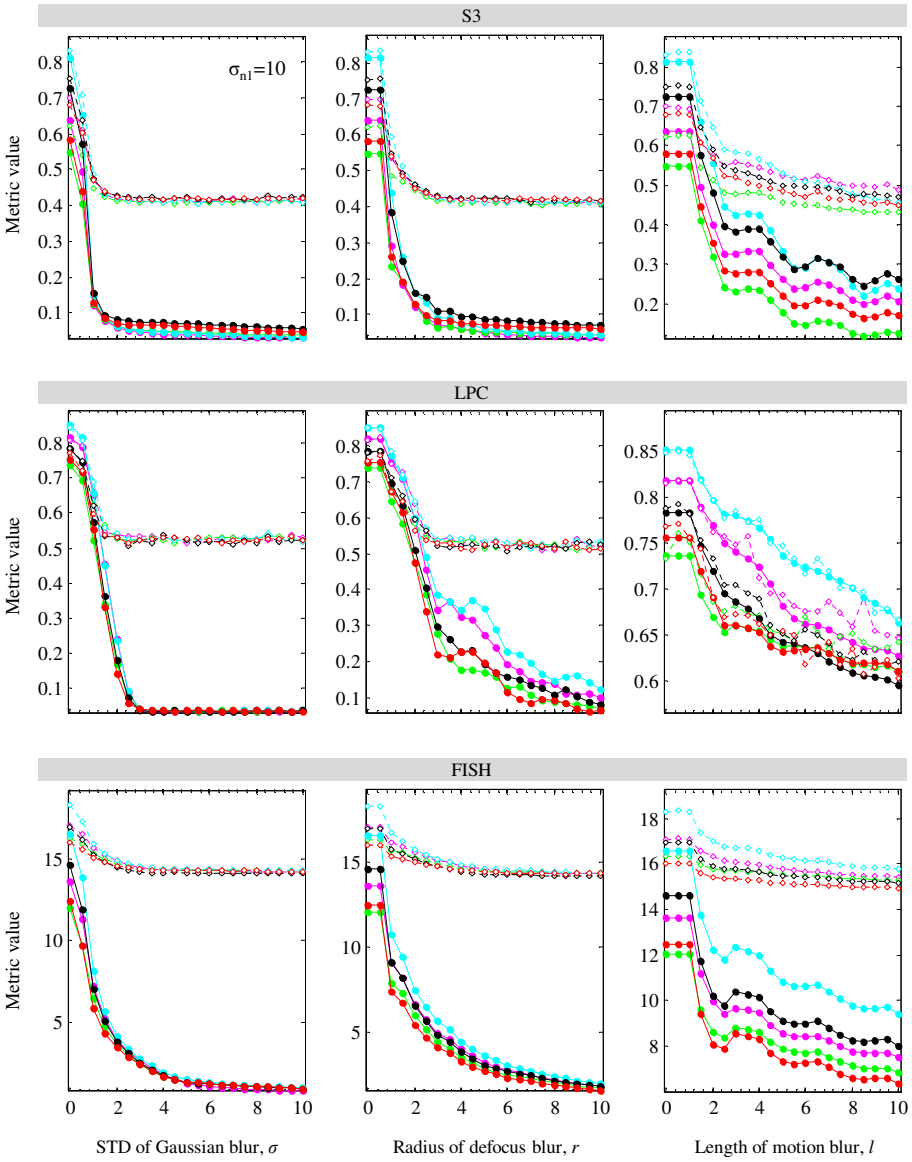


Figure 5.34 (Continued)

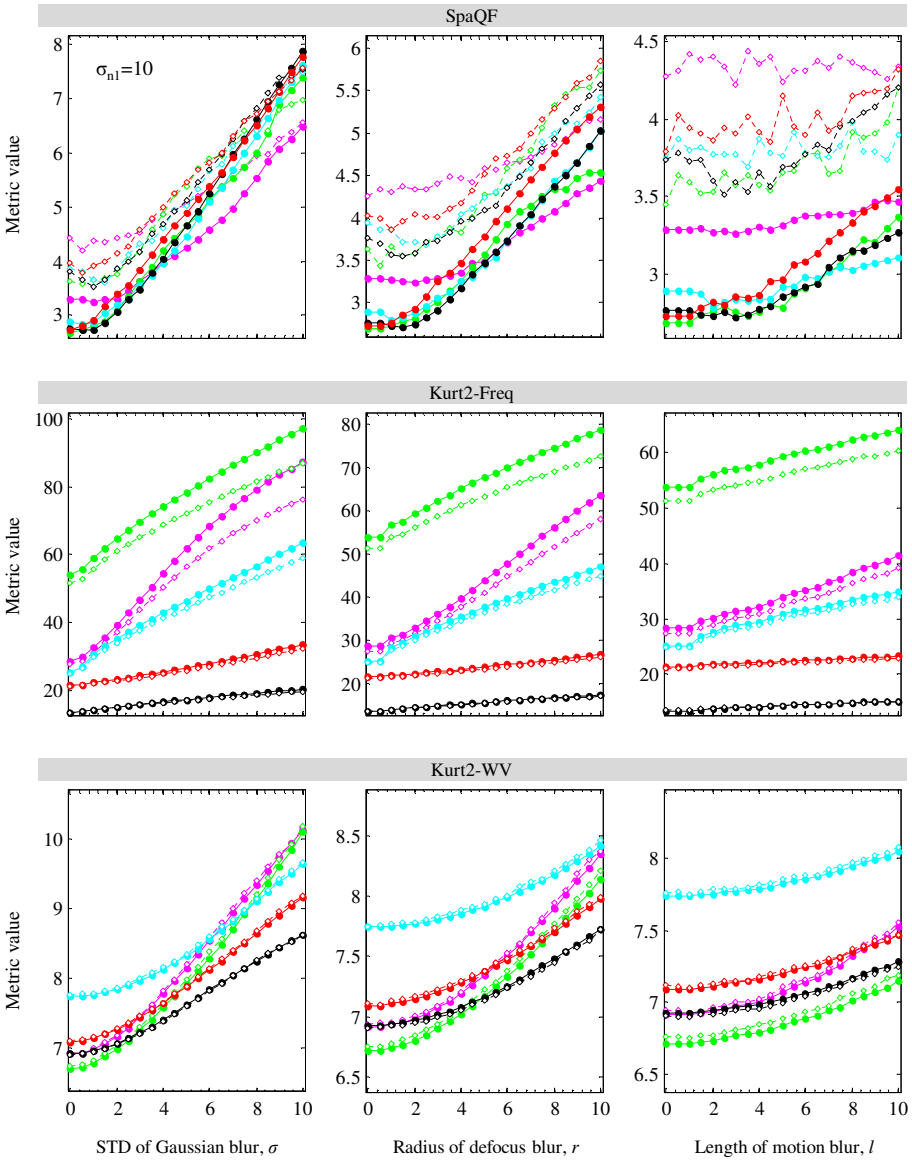


Figure 5.34 (Continued)

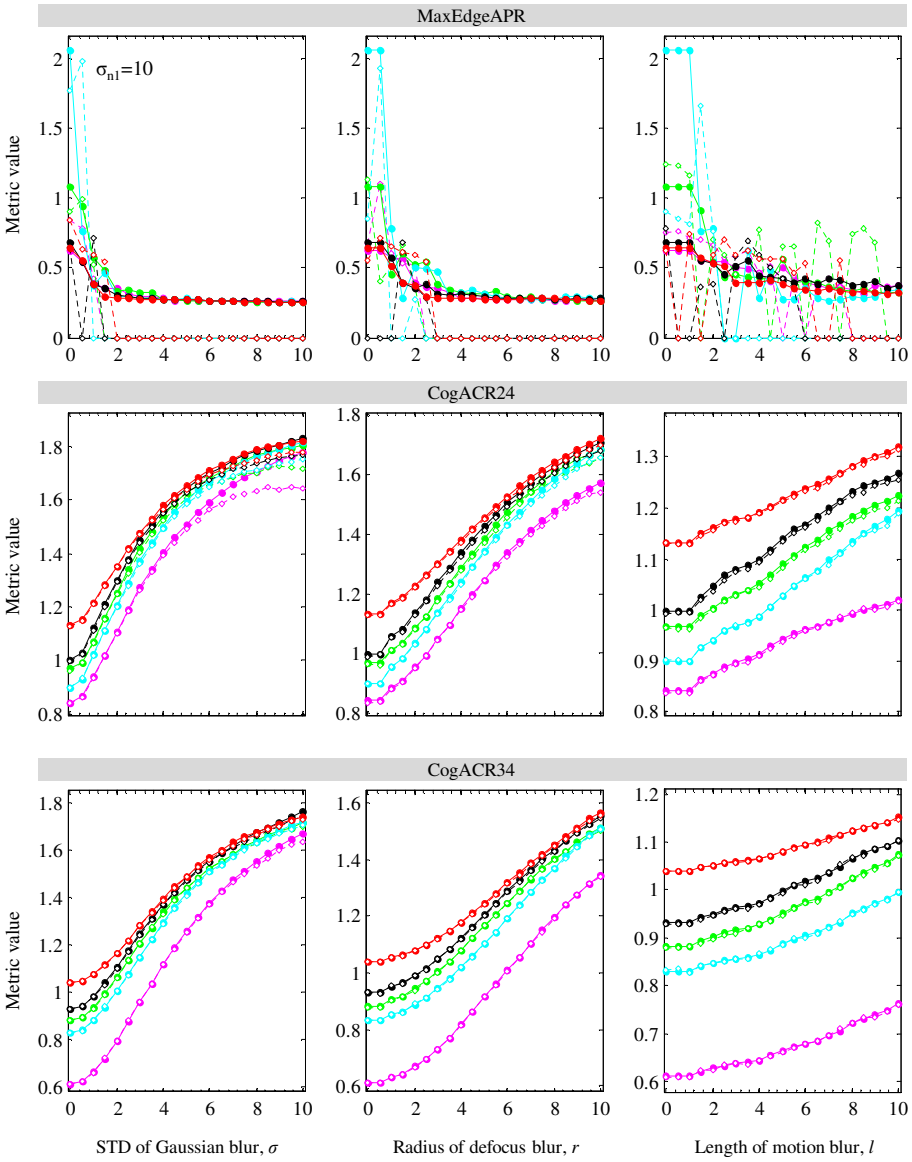


Figure 5.34 (Continued)

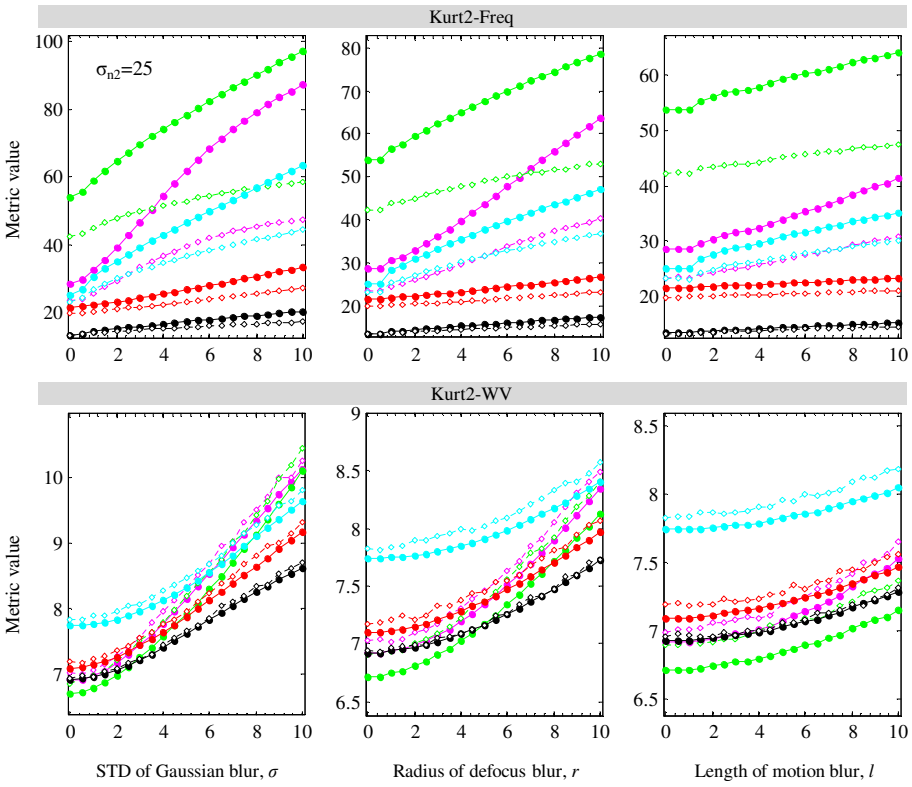


Figure 5.35: Performance of (top) Kurt2-Freq and (bottom) Kurt2-WV NR blur measures for images corrupted with noise of $\sigma_{n2} = 25$. All other descriptions are the same as in Figure 5.34.

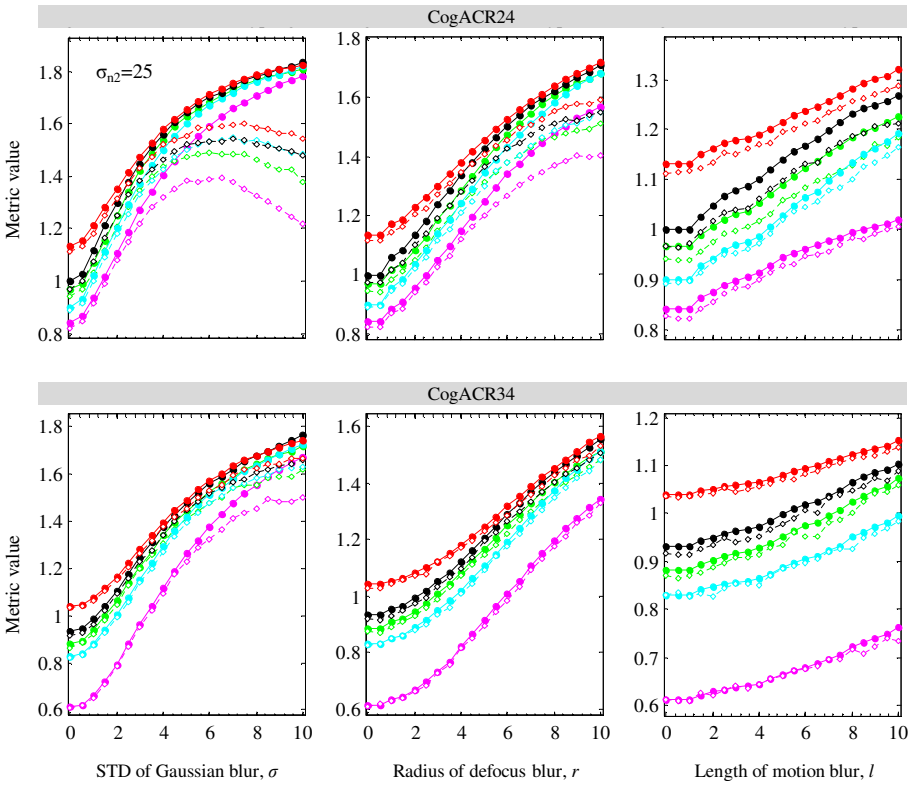


Figure 5.36: Performance of the proposed (top) CogACR24 and (bottom) CogACR34 NR blur measures for images corrupted with noise of $\sigma_{n2} = 25$. All other descriptions are the same as in Figure 5.34.

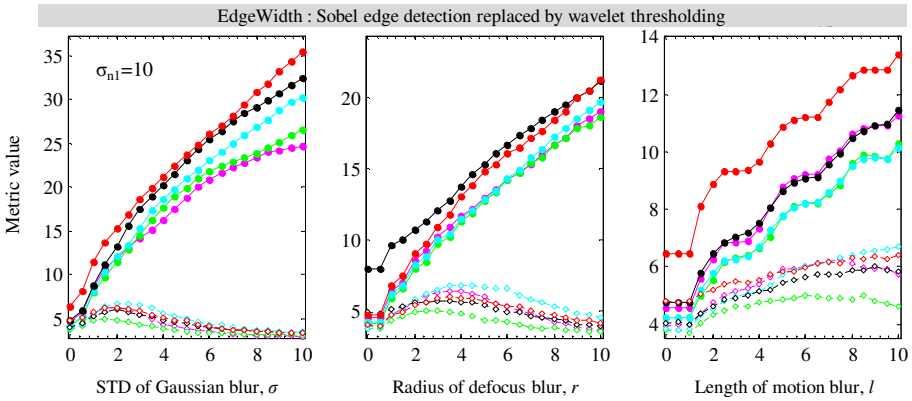


Figure 5.37: Evaluation of the effects of edge detection. Performance of the EdgeWidth measure when the original Sobel detector is replaced by the proposed thresholding of wavelet inter-scale products. All other descriptions are the same as in Figure 5.34.

together with noise) for which also the human IQ ratings are available; the details are described in Section 5.8.1.3. We assess the same set of state-of-the-art NR blur measures as in the previous sections but now with regards to two different aspects: (1) the agreement between the measure and the true parameters of GBlur and (2) the ability of the measure to predict humans. As the indicator of the performance, we compute the SROCC between the NR measure scores and the estimated parameter, either the true amount of blur or the MOS values of humans (we refer to Section 5.8.1.2 for more information about MOS). For further information about the SROCC, we refer to Section 5.8.4.

Table 5.3 summarizes the SROCC values computed between the measure and the true amount of introduced GBlur. Note that the results are presented for several different subsets of images, created according to the level of image noise. This is done in order to allow more detailed analysis of the effects of noise. The images are arranged in four groups: TestGWN0, noise-free images only (15 REFs plus 45 with GBlur of $\sigma_b \in \{3.2, 3.9, 4.6\}$); TestGWN1, TestREF together with the 45 images distorted with $\sigma_{n1} = 11.5$; TestGWN2, TestGWN1 together with the 45 images distorted with $\sigma_{n2} = 22.8$; and TestGWN3, TestGWN2 together with the 45 images distorted with $\sigma_{n3} = 45.6$. According to the SROCC values, the SpaQF measure correlates best with the true parameters of GBlur. Remember also from Section 5.7 that this measure is primarily designed for GBlur identification. The second and the third best are the proposed measures, respectively, CogACR24 and CogACR34. Note that the levels of blur in LIVE2BlurNoise database are relatively small ($\sigma < 5$) compared to those in our previous tests ($\sigma \leq 10$) and thus CogACR24 measure is little affected by noise; this allows CogACR24 to outperform CogACR34. Considering the effects of noise, only the highest level of noise ($\sigma_{n3} = 45.6$ which occurs in TestGWN3) seems to

have an affect on the three selected measures. However, the increasing level of noise clearly affects some other measures, *e.g.*, CPBD, JNBM, FISH, S3, and LPS. This is to be expected based on the previous experiments for different types of blur.

Lastly, Table 5.4 quantifies the correlation between the measure and the human MOS values. The SpqQF measure remains the one with highest SROCC values among the tested measures. In fact, it achieves a rather high correlation of SROCC =0.928 for noise-free images (image group TestGWN0), stays nearly unaffected by the lowest considered level of noise ($\sigma_{n1} = 11.5$ for TestGWN1), and then drops down to SROCC =0.831 when the larger noise levels are also included ($\sigma_{n2} = 22.8$ and $\sigma_{n3} = 45.6$ for TestGWN3). The FISH measure achieves nearly the same high level correlation for TestGWN0, SROCC =0.917. However, the correlation rapidly drops in the presence of image noise (SROCC =0.659 for TestGWN1, and SROCC =0.055 for TestGWN3). Similar is true for a few other measures CPBD, JNBM, S3, and LPC. Remember from Section 5.7 and Table 5.1 that the formed three of these methods are aimed by design to predict humans. Overall, the proposed two measures rank the second and the third best, after the SpaQF. Thus, the overall ranking of the measures is the same as for the correlation with the true level of GBlur. What differs is the absolute range of SROCC values; these are now considerably lower for CogACR measures (SROCC =0.853 for TestGWN0) compared to the SpaQF (SROCC =0.928 for TestGWN0).

In order to better understand the observed differences between the performance of the SpaQF and of the proposed two measures, we plot in Figure 5.38 the measures values for all images of LIVE2BlurNoise. As can be seen from the plots, the values of the SpaQF measure are nicely grouped for the lower levels of blur and they disperse at the higher blurs. In contrast, the CogACR measures are more dispersed at lower blurs and they gather closer as the blur increases. This "complementary" behavior explains the very similar SROCC ranking of the three measure in terms of their correlation with the true amount of blur. Humans, on the other hand, rate the IQ in a nonlinear fashion, *i.e.*, they are more sensitive to the changes of image blurriness at small levels of blur then to the changes within the high range of blur. Accordingly, the measure which agrees better with humans at the low levels of blur can be expected to achieve higher overall correlation with humans. This offers an explanation for the earlier observation that the SROCC value of SpaQF and that of CogACR measure differ more in Table 5.4 than in Table 5.3.

Table 5.3: SROCC computed between the NR measure and the true amount of introduced GBlur for the LIVE2BlurNoise image database. The images are grouped based on the amount of added noise and the SROCC is computed for each group (represented by rows in the table, from top to bottom): TestGWN0 – noise-free images only (15 REFs plus 45 with GBlur of $\sigma_b \in \{3.2, 3.9, 4.6\}$); TestGWN1 – TestREF together with the 45 images distorted with $\sigma_{n1} = 11.5$; TestGWN2 – TestGWN1 together with the 45 images distorted with $\sigma_{n2} = 22.8$; and TestGWN3 – TestGWN2 together with the 45 images distorted with $\sigma_{n3} = 45.6$. For details regarding the LIVE2BlurNoise image database, see Section 5.8.1.3.

	CPBD	JNBM	Kurt2-WV	EdgeWidth	Kurt2-Freq	CogACR24	CogACR34	FISH	EdgeAPR	SpaQF	S3	LPC
TestGWN0	0.940	0.887	0.199	0.026	0.233	0.935	0.913	0.964	0.179	0.950	0.848	0.935
TestGWN1	0.343	0.699	0.199	0.135	0.234	0.935	0.913	0.737	0.170	0.949	0.713	0.838
TestGWN2	0.094	0.442	0.198	0.228	0.230	0.933	0.913	0.485	0.156	0.949	0.642	0.734
TestGWN3	0.030	0.268	0.196	0.280	0.226	0.928	0.913	0.367	0.114	0.946	0.604	0.683

Table 5.4: SROCC computed between the NR measure and the human MOS values. Each row corresponds to a different group of images from the LIVE2BlurNoise image database, the same as in Table 5.4.

	CPBD	JNBM	Kurt2-WV	EdgeWidth	Kurt2-Freq	CogACR24	CogACR34	FISH	EdgeAPR	SpaQF	S3	LPC
TestGWN0	0.871	0.816	0.317	0.059	0.358	0.853	0.830	0.917	0.138	0.928	0.834	0.865
TestGWN1	0.257	0.612	0.283	0.258	0.339	0.859	0.839	0.659	0.163	0.923	0.667	0.766
TestGWN2	0.093	0.248	0.285	0.432	0.340	0.817	0.802	0.259	0.255	0.878	0.490	0.570
TestGWN3	0.274	0.004	0.279	0.530	0.316	0.782	0.777	0.055	0.302	0.831	0.374	0.452

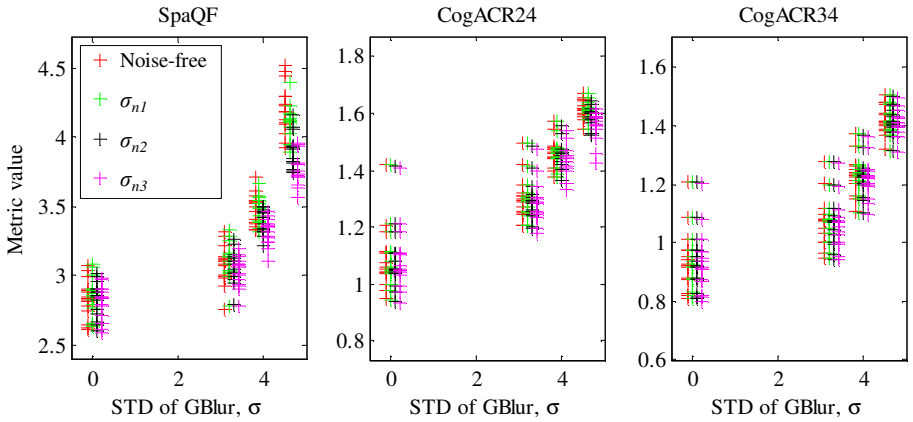


Figure 5.38: Performance of the three best NR blur measures according to the results in Table 5.4 and Table 5.3 (from left to right): SpaQF, CogACR24, and CogACR34. The measure values correspond to the LIVE2BlurNoise database (all images included). Different colors represent different levels of image noise.

Note at the same time that CogACR is sensitive to image content in the way which intuitively complies to the sensitivity of the human visual system, higher sensitivity to small distortions in high frequency image content compared to the low frequency one. This suggests potential for using CogACR to build an algorithm for objective evaluation of the perceptual quality of images. Going in this direction, it seems worthwhile to join the emerging trend of adding saliency in the objective IQA [Ninassi et al., 2007, Jänicke and Chen, 2010, Engelke et al., 2011, Liu et al., 2013]. Moving in that direction, it would be of interest to explore a couple of alternatives; for example, restricting the area of acting of the proposed measure to the the salient regions of image (rather than on the whole image area as we do it now) or using the saliency information as a weighting factor for the individual ACR coefficients. This is for future research to examine.

5.8.6 Real-time performance for high-definition data

In order to achieve real-time performance, the implementation of the proposed CogACR measure has been adapted to take into account hardware specifications of a commercially available IBM Cell BE multi-core microprocessor architecture (3 out of the 8 SPE cores were used for real-time performance). The results suggest that, although computationally demanding (based on the non-decimated wavelet transform), the proposed measure can be efficiently implemented on a commercial platform and achieve real-time performance for high-definition (HD) input. Moreover, with this implementation, the CogACR method has been incorporated in an existing video quality

assessment platform [Papp et al., 2009] and tested with a commercial HD Set Top Box (STB); in particular, the Sky+HD box from British Sky Broadcasting.

This work was performed during the year 2009 in collaboration with Nemanja Lukić and Prof. Miodrag Temerinac from Novi Sad University, Serbia [Lukić et al., 2010].

5.9 Conclusion

The work presented in this chapter addressed the problem of blur identification focusing primarily on the NR scenario of IQA where the distortion-free image is not available. We started with defining a method for edge detection which is highly robust to both blur and noise. Further analysis towards assessment of the image blur considered only the selected edge pixels. The first main contribution of this work is the novel NR blur measure named CogACR. In comparison to a number of state-of-the-art NR blur measures, our experimental results suggested several major advantages of the proposed method: (1) CogACR behaved monotonically for all three tested types of blur (GBlur, DBlur, and MBlur); the same was found for only two other considered measures (Kurt2-Freq and Kurt2-WV), (2) when tested for five different image contents in the presence of both blur and noise, CogACR34 outperformed all tested measures in terms of robustness to noise, it remained nearly unaffected by a very high level of noise, (3) when tested for the LIVE2BlurNoise publicly available image database with multiple distortion (BGlur and noise together), CogACR performed second best (after SpaQF which is specifically designed for GBlur) both in terms of the SROCC correlation with the true blur levels and in terms of the SROCC correlation with the human ratings of IQ.

Furthermore, we examined the effects of image content on the performance of blur measures. As a descriptor of image content, we proposed using the histogram of ACR values (HistACR) corresponding to the dominant edges in the image. Moreover, we proposed a novel HistACR-based measure of image similarity. While existing similarity measures are often context-based, our technique quantifies the similarity of edges in the images. When tested for the LIVE1Blur database of images with GBlur, this method could successfully identify the images which behave similarly in the presence of varying levels of blur, not only according to the CogACR measure but, more importantly, also according to the human ratings of the IQ.

Outside of the domain of multimedia images, the proposed CogACR method has been successfully used as a tool for blur estimation in medical video in the context of two consortium projects, both financially supported by iMinds: the “Telesurgery” project, which assessed the quality of laparoscopic surgery videos, and the ongoing “Ultra Wide Context Aware Imaging” (PANORAMA) project evaluating the quality of x-ray coronary angiographic image sequences. That work was coordinated by Asli Kumcu (Department of Telecommunications and Information Processing, Ghent University, Belgium).

Finally, in the scope of our investigations in collaboration with the Faculty of Technical Sciences, Novi Sad, Serbia the CogACR measure has been efficiently implemented on a commercially available processor and integrated in an existing video quality assessment platform where it has been tested with a commercially available HD Set Top Box.

The contributions reported in this chapter resulted in two international conference publications [Ilić et al., 2009, Platiša et al., 2011j] and two conference talks [Platiša et al., 2010d,Platiša et al., 2011i]. Another conference publication is a result of collaboration with Nemanja Lukić and prof. Miodrag Temerinac from Novi Sad University, Serbia which led to a real-time implementation of the proposed blur measure [Lukić et al., 2010]. A journal paper is currently in preparation [Platiša and Pižurica, 2014].

6

Quality of appearance

While the previous chapters focused on estimating the effects of image degradation on the (technical or task-based) image quality, in this chapter we assess the *effects of an art painting technique* on the quality of appearance of the painted objects. In particular, we study the problem of evaluating the appearance of pearls in the images of art paintings for the purpose of developing tools for art historical analysis. To that end, we develop numerical methods which capture the attributes of the pearl's appearance, *e.g.*, the appearance of the smoothness of the pearl's surface. Our analysis is based on the so-called image spatiograms which extend the conventional histograms with the spatial information; a key factor in the analysis of appearance. The principal case study for this work is the world-famous 15th-century polyptych *Ghent Altarpiece*, located in the Saint Bavo Cathedral in Ghent, which is currently undergoing a major five year restoration project.

Preface

In his 1435 treatise on the theory of painting *De pictura* ("On Painting"), Leon Battista Alberti remarks that "gems and all precious things of that kind become much more precious by the painter's hand" [Sinisgalli, 2006]. Indeed, artists have been attracted to the beauty of precious stones and challenged to depict jewelery in their paintings for ages. As comprehensively reviewed by [Autin et al., 1999] in the recent *Jewels in Painting*, the first pearls appear in the paintings of 15th-century artists. These include the *Portrait of a Young Lady* by Petrus Christus and the *Portrait of Queen Margaret of Denmark* by Hugo Van der Goes, in Flanders, and the *Portrait of Simonetta Vespucci* by Piero di Cosimo and the *Portrait of Battista Storza* by Piero della Francesca, in Italy. Famous later examples include, of course, the depictions in Johannes Vermeer's *Girl with the Pearl Earring* and the *Turkish Bath* by Jean-Auguste-Dominique Ingres.

The way an artist paints pearls reflects their ability to observe nature, and in some cases, such as Jan Van Eyck in the *Ghent Altarpiece* masterpiece, their acquaintance with contemporary optical theory [De Mey, 2008]. The painterly execution may also

be considered as an idiosyncratic marker or an individual characteristic useful in distinguishing hands. In this context, we resort to the framework of digital image analysis to study the painterly technique of an artist. Specifically, we study the problem of the two-dimensional (2D) representation of reality, and propose a method which aspires to create a tool for art historical attribution.[1]

6.1 Introduction

Traditionally, analysis and interpretation of paintings is driven by researchers with primary expertise in art history. Nevertheless, recent years have evidenced great advantages of "collaborative art history" [Silver, 2006] - adjacent disciplines working together to provide different perspectives and deeper insights into the research questions.

One current trend among cultural heritage institutions is photographing their works of art for a variety of applications [Farnand et al., 2009], ranging from those primarily commercial such as promotional websites, exhibit catalogues and other printed materials for sale in museum shops, to the research and teaching oriented ones such as digitized records for the art conservation and digital archives of cultural heritage.[2]

In this chapter, we attempt to assist art historians in studying the *quality of appearance* of the painted jewels, such as pearls and beads.[4] Thereby, we take the general framework of this thesis, the image quality assessment, beyond its conventional boundaries of beauty or utility. Rather, we explore the attributes of appearance of objects in the images (here, of pearl-like objects in paintings) and seek to develop numerical methods which capture those attributes.

Our proposed techniques are based on the so-called *spatiograms* [Birchfield and Rangarajan, 2005] of images which extend the concept of histograms to the spatial domain. Knowing that surface reflectance is among the most notable characteristics of jewels in paintings, it was essential to have spatial information involved in the analysis of pearl images. The contribution of our work starts with a demonstration of the ability of an existing spatiogram similarity measure [Conaire et al., 2007] to quantify the overall similarity between pearl images. At the same time, we point to its major weakness for the analysis of painted objects - the lack of ability to inform about specific aspects of object appearance. Moreover, we propose a method for visualizing the multidimensional spatiogram data; the problem which has not been addressed before.

Secondly, we introduce a method for matching spatiograms of different images

[1] It is important to note that the problem of the *two-dimensional representation of reality* studied here differs from the problem of how *optical "reality"* is rendered in the panels studied elsewhere [Stork, 2006, Stork and Johnson, 2006, Stork and Duarte, 2007].

[2] Another possibly expanding application domain for the digitized artworks might be the online art trade, triggered by the very recent Amazon Art, a fine-arts and collectibles category launched in August 2013 at the world's largest online retailer Amazon.[3]

[4] While the discussion in this chapter is centered on images of pearls and beads, it largely applies also to the quality of appearance of other small spherically-shaped objects in paintings.

and use it in our explorative analysis of the dominant factors of the appearance of pearl-like objects. More generally, this technique could be extended to enable virtual style manipulations (outside of an art historical context). As an example, we could think of creating virtual copies of the very famous *Mona Lisa* by Leonardo da Vinci such that each copy adopts the style of another master painter to depict the ever intriguing lips in the portrait and then studying the (aesthetic) effects of the different painterly styles.

Thirdly, we propose a set of novel measures based on the spatiograms of pearl images which quantify numerically a set of perceptually relevant object features; mainly, the appearance of surface smoothness and several aspects regarding object symmetry. As we will see later in the chapter, the proposed techniques can be used in multiple manners, including numerical quantification of the visually observed image features and the degree of realism of the visual appearance in the painting, characterization of the specific properties of rendering of different materials by an artist, or detecting copies of the artworks.

To test the performance of our proposed techniques we use both the images of painted pearls and those of real ones. For the pearls in paintings, we look at the *Ghent Altarpiece*, both the pearls painted by the original masters, the Van Eyck brothers (1432), and those of their copyists, Jef Van der Veken (1945) and Charlotte Caspers (2010). In addition, we consider the pearls painted by Hans Memling in his *Portrait of Maria Maddalena Baroncelli* (1470). The digital images of the artwork by Van Eyck are based on photograph negatives from the late Alfons Dierick (c-04, h-16, 40-15) made available for research purposes to Ghent University. We thank Saint Bavo cathedral, Lukas-Art in Flanders, and the Dierickfonds for permission to use these materials for the research reported in this chapter and in the related publications arising from this work (listed in Section 6.5).

We became involved in this research on the initiative of Prof. Ingrid Daubechies (Mathematics Department, Duke University, USA) who put us in contact with Prof. Marc de Mey (Royal Flemish Academy of Belgium for Science and the Arts (KVAB), Belgium), Prof. Maximiliaan Martens, Dr. Annick Born, and Dr. Emile Gezels (Department of Art, Music and Theatre Sciences, Ghent University, Belgium). In addition, this research involved collaboration with Prof. Ann Dooms and Bruno Cornelis (Department of Electronics and Informatics, Free University of Brussels, Belgium).

The rest of this chapter is organized as follows. In Section 6.2, we first discuss in more detail the necessity for spatially-aware techniques in the case of analysis of object appearance; this motivated our choice for spatiograms over histograms. Afterwords, the concept of spatiograms is outlined and the existing spatiogram-based measures are reviewed. The end of the section draws attention to the main drawbacks of the existing similarity measures for the application at hand and suggests the expected benefits from the new approach. The investigations from our work and the proposed techniques are presented in Section 6.3. Our experimental study results are presented and discussed in Section 6.4. Finally, some concluding remarks and possi-

ble directions for future research are given in Section 6.5.

6.2 Digital images of painted pearls

To study the images of pearls, it is essential to quantify the distinctive features that evoke the visual impressions of the pearl-like luster and sheen. Painted pearls, for example those in the *Ghent Altarpiece*, are often characterized by a blurry *highlight* as a reflection image of the light source and a fine glowing line at the other side indicating the delicate *sheen* emanating from the surface [De Mey, 2008]. For illustration, see the example pearl images in column three of Figure 6.1 and column four of Figure 6.4 as well as the pearls in Figure 6.12 (a) and (b). Another type of pearl-like objects which frequently appear in the *Ghent Altarpiece* are beads. In contrast to the pearls which have specular (mirror-like) reflection, the highlight areas of beads are usually painted smaller and with fine sharp edges, giving an impression of matte surfaces with the diffuse reflection property; for illustration, see the beads in column two of Figure 6.1 and columns one to three of Figure 6.4.

A simple and effective way to statistically characterize the distribution of pixel values in an image is to count and tabulate the number of occurrences of a given value (or a given interval of values). This representation is called the *histogram*. Thus, a histogram tells us which fraction of image pixels belongs to which range of brightness values. Digital image histograms can be used to discriminate between different materials in the scene: glass, wood, metal, and other. Figure 6.1 shows histograms of three spherical objects from the *Ghent Altarpiece*, each from a different material: a metal ball, a glass bead and a pearl. The corresponding three histograms clearly differ. However, histograms describe only global distribution of pixel values in the image without capturing their spatial relations. In other words, the histogram of an image informs only of the relative proportion of each brightness value but not of their spatial positioning and spreading. It remains unclear in which part(s) of the image a particular pixel value appears; is it concentrated in a smaller region or is it scattered all over the image? Take as an example the two pearls from Figure 6.2. Their histograms are identical while their visual appearance is obviously different. The highlight in the top pearl is elliptical and quite sharply delineated, while the one in the bottom pearl is irregular and fades away smoothly. The inability of the histograms to distinguish between such important features of the painted objects motivated us to search for more advanced statistical characterization methods, such as *spatiograms* [Birchfield and Rangarajan, 2005].

6.2.1 Digital image spatiograms

The image spatiogram is an extension of the histogram representation such that certain spatial features of the image data are taken into account, next to the global distribution of the pixel values. Not only the fraction of pixels with values within a certain range is

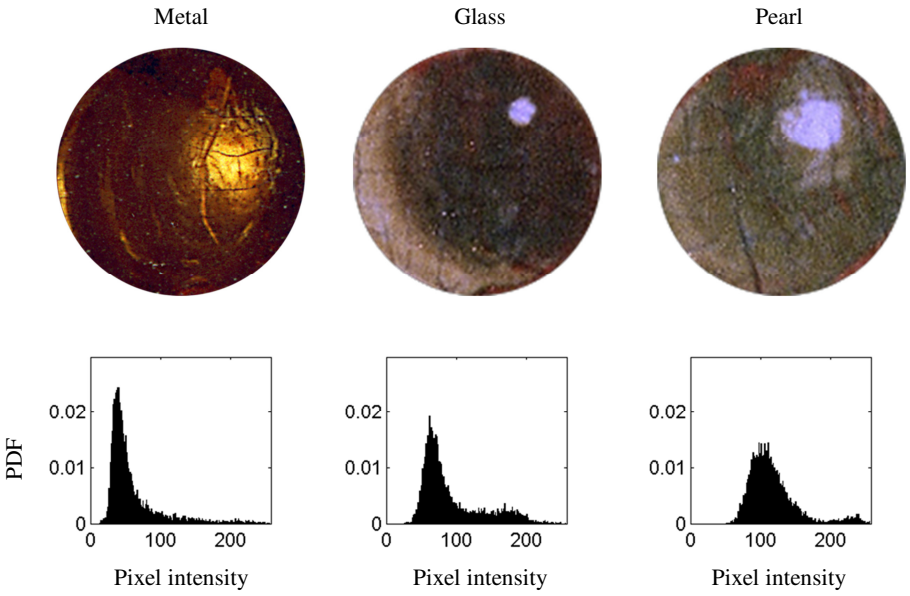


Figure 6.1: Normalized histograms for painted objects made of three different materials (from the Ghent Altarpiece).

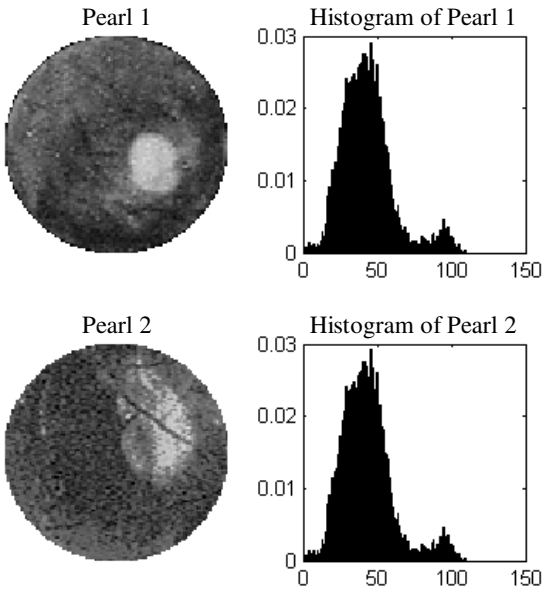


Figure 6.2: Two visually different pearls which have identical histograms.

captured, but their spatial positioning, too. The concept of a spatial histogram, or spatiogram, was first introduced by [Birchfield and Rangarajan, 2005] who were aiming to improve the performance of object tracking systems (hence the term "spatiogram trackers"). They formulate a spatiogram as a generalization of a histogram, that is, a histogram with second-order spatial moments, where a histogram is a zeroth-order spatiogram.

Given a 2D grayscale image of N pixels and a set of B bins identified by index $b \in \{1, ..., B\}$, we can write a spatiogram for the current bin as a triplet $\mathbf{S}_b = (c_b, \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$. Here, the triplet elements represent, respectively:

- the normalized histogram bin count;
- the spatial centroid of the pixels in a given bin (mean x- and y-coordinates), $\boldsymbol{\mu}_b = (\bar{x}_b\ \bar{y}_b)$;
- the matrix of spatial covariances (variation in x- and y-coordinates), $\boldsymbol{\Sigma}_b = \begin{pmatrix} q_b(x,x) & q_b(x,y) \\ q_b(y,x) & q_b(y,y) \end{pmatrix}$.

Finally, assuming the columns of $\boldsymbol{\Sigma}_b$ are vertically stacked into a single column vector, the spatiogram for all bins can be written as a triplet $\mathbf{S} = (\boldsymbol{c}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ where each element is comprised of B columns.

If we let the vector $\mathbf{p}_n = (x_n\ y_n)^T$ denote the spatial position of the current pixel n, $n \in \{1, ..., N\}$, the spatiogram triplet for bin b is computed as follows:

$$c_b = \eta \sum_{n=1}^{N} \delta_{nb}, \quad \delta_{nb} = \begin{cases} 1, & \text{if pixel } n \text{ belongs to bin } b \\ 0, & \text{otherwise} \end{cases}, \tag{6.1}$$

$$\boldsymbol{\mu}_b = \frac{1}{\sum_{j=1}^{N} \delta_{jb}} \sum_{i=1}^{N} \mathbf{p}_i \delta_{ib}, \tag{6.2}$$

$$\boldsymbol{\Sigma}_b = \frac{1}{\sum_{j=1}^{N} \delta_{jb}} \sum_{i=1}^{N} (\mathbf{p}_i - \boldsymbol{\mu}_b)(\mathbf{p}_i - \boldsymbol{\mu}_b)^T \delta_{ib}. \tag{6.3}$$

The normalising constant η in Eq. (6.1) is chosen such that $\sum_{b=1}^{B} c_b = 1$. Formally, for bins with $c_b = 0$ also the values of $\boldsymbol{\mu}_b$ and $\boldsymbol{\Sigma}_b$ are set to zero (not of interest). Finally, to enable comparison between images of different sizes, all spatial coordinates need to be normalized to the same range; in our experiments [-1,1].

Figure 6.3 shows a constructed example of two simplified ball-shaped image objects which illustrates a situation where spatiograms are superior to histograms. Note that the two histograms are exactly the same (for the two objects, the counts of pixels with specific intensity values are the same) and thus not able to differentiate between

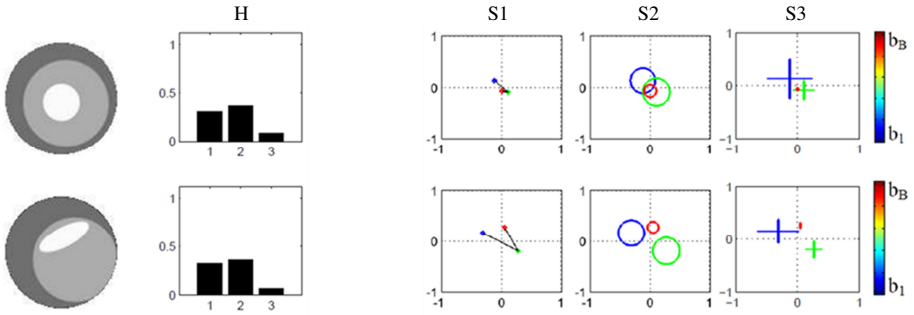


Figure 6.3: Pearl histogram versus pearl spatiogram. Left to right: input images, histograms, the three spatiogram plots S1, S2 and S3 (see text).

the objects. In contrast, the plots of the two spatiograms (graphs S1, S2, S3) are obviously different (for the two objects, the pixels with the same intensity values are spatially arranged in different ways). The details of the spatiogram visualization are elaborated in Section 6.3.1. In the case of more complex objects, such as the painted pearls and beads from Figure 6.4, the high-dimensional spatiogram representation becomes less intuitive for interpretation and also quite complex for visual inspection. In those real world applications it is handy to have numerical tools at hand that are capable of characterizing the spatiograms and preferably also of quantifying specific features of spatiograms. We first review the tools for comparing two spatiograms.

6.2.2 Existing spatiogram similarity measures

Several methods have been proposed to compare two spatiograms. The original method by [Birchfield and Rangarajan, 2005] suffered from a major drawback of failing to ensure that comparing a spatiogram to itself produces a constant value (*e.g.* if two images of different content are compared to themselves, the similarity scores may differ). Later works by [Ulges et al., 2006, Conaire et al., 2007, Gong et al., 2009b, Yao et al., 2011] successfully resolved this problem and at the same time attempted to improve the discriminative power of the measures. We describe next the basic elements common to all mentioned methods and specify the one used in our study; further details are beyond the scope of our work and the interested reader is referred to the original papers.

The similarity between two spatiograms is computed as the weighted sum of the histogram bin similarities. If we let $\mathbf{S}_1 = (\boldsymbol{c}_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{S}_2 = (\boldsymbol{c}_2, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ denote two different spatiograms, each with B bins, the similarity ρ between $\mathbf{S}_1$ and $\mathbf{S}_2$ can be written as

$$\rho(\mathbf{S}_1, \mathbf{S}_2) = \sum_{b=1}^{B} \phi_b(\boldsymbol{c}_{b1}, \boldsymbol{c}_{b2}) \psi_b(\mathbf{S}_1, \mathbf{S}_2), \tag{6.4}$$

where ϕ_b is the similarity between the b-th bin histograms (often computed as his-

togram intersection or Bhattacharyya coefficient) and ψ_b is the spatial similarity for those same bins. The existing similarity measures differ from each other in the factor ψ_b they apply to the histogram bin similarities ϕ_b.

Commonly, spatial positions $\mathbf{p}$ of pixels from the same bin b are assumed to have Gaussian distribution described by the intra-bin statistics of $\boldsymbol{\mu}_b$ and $\boldsymbol{\Sigma}_b$, that is $\mathbf{p} \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$, or

$$f(\mathbf{p}|b) = \frac{1}{2\pi|\boldsymbol{\Sigma}_b|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{p}-\boldsymbol{\mu}_b)^T {\boldsymbol{\Sigma}_b}^{-1}(\mathbf{p}-\boldsymbol{\mu}_b)\right). \tag{6.5}$$

The Gaussian model is obviously a substantial simplification (it holds only when the pixels from the same bin are grouped in a single blob, such as, for example, the pixels from the highlight area of a pearl image). Nevertheless, it brings the problem of measuring spatial similarity down to comparing two Gaussian probability distributions where multiple existing tools become available, for example, Mahalanobis distance of the means [Birchfield and Rangarajan, 2005], Jensen-Shannon divergence [Ulges et al., 2006], Bhattacharyya distance [Conaire et al., 2007], symmetric Kullback-/Leibler distance [Yao et al., 2011], as well as some more recent methods such as Lie group distance defined specifically for image representation purposes [Gong et al., 2009a, Gong et al., 2009b].

One of the more studied spatiogram similarity measures, which we also use in our work, is the method proposed by [Conaire et al., 2007]. They use the Bhattacharyya coefficient to compare the two model Gaussians, $\mathcal{N}(\boldsymbol{\mu}_{b1}, \boldsymbol{\Sigma}_{b1})$ and $\mathcal{N}(\boldsymbol{\mu}_{b2}, \boldsymbol{\Sigma}_{b2})$, formulating the final similarity measure as follows:

$$\phi_b(\boldsymbol{c}_{b1}, \boldsymbol{c}_{b2}) = \sqrt{\boldsymbol{c}_{b1}\boldsymbol{c}_{b2}}, \tag{6.6}$$

$$\psi_b(\mathbf{S}_1, \mathbf{S}_2) = \alpha_b \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_{b1}-\boldsymbol{\mu}_{b2})^T {\hat{\boldsymbol{\Sigma}}_b}^{-1}(\boldsymbol{\mu}_{b1}-\boldsymbol{\mu}_{b2})\right), \tag{6.7}$$

where $\hat{\boldsymbol{\Sigma}}_b = 2(\boldsymbol{\Sigma}_{b1}+\boldsymbol{\Sigma}_{b2})$ and $\alpha_b = \frac{4|\boldsymbol{\Sigma}_{b1}\boldsymbol{\Sigma}_{b2}|^{\frac{1}{4}}}{|\hat{\boldsymbol{\Sigma}}_b|^{\frac{1}{2}}}$. The similarity score ρ between two spatiograms ranges from 0 to 1, where the maximum similarity score $\rho = 1$ is reached only when the spatiogram is compared to itself. We discuss this measure further in Section 6.4.2.

6.2.3 What is still missing?

While they slightly differ in the specifics, all existing similarity measures compare spatiograms at a global level and result in a single number indicating the overall similarity between the two spatiograms, that is, between their corresponding images. Unfortunately, even if they were able to provide a perfect measure of the overall pearl similarity, these measures are not sufficient for our purpose. What we are looking for is a way to quantify distinctive spatial features which are indicative of the visual

appearance and, moreover, of some specific attributes of the appearance. To put it back in the terms of spatiograms, we want to know why in certain cases ρ is small or big and to express the contributing factors explicitly. We want to learn about the specific aspects of different painterly representations that contribute to smaller or larger similarity scores of their spatiograms.

As we will demonstrate in our experiments reported in Section 6.4.2 and Section 6.4.3, the similarity score ρ is useful for our analysis, but definitely not sufficient. Therefore, in the following Section 6.3, we set to explore in more depth the relationship between the visual appearance of pearls and their spatiograms, and based on those observations, we define additional spatiogram-based measures for our analysis. Eventually, we conduct a simple experiment with human observers and assess in Section 6.4.5 the potential of our method to practically characterize the appearance of pearl images.

6.3 Quality of appearance of pearls in the images

In this section we introduce the novel mathematical descriptors of the appearance of pearls and pearl-like objects in the images. First, we propose three types of plots as a means of visualizing the high dimensional spatiogram data. These plots are used as tools for the first step of our research process: exploratory visual inspection of the spatograms which correspond to the images of visually similar pearls. The goal here is to identify spatiogram features which are most discriminative of specific attributes of object's visual appearance (*e.g.* the impression of uniformity of the pearl's surface, the impression of pearl's reflectance, etc.). In addition, we propose a method for transforming an image by matching its spatiogram to that of another image. This technique allows for the reverse analysis, *i.e.*, analysis of how the appearance of a given image is changing with its spatiogram – the next step of our investigations towards quantifying the appearance. Finally, relying on the previous observations, we define four new spatiogram-based measures which could be used to characterize the attributes of appearance of pearls in the images. The details are presented next.

6.3.1 Visualization of spatiograms

Given the goal of our research here - developing mathematical methods to objectively characterize the visual appearance of pearls in the paintings, we choose to begin our problem analysis by visual inspection of pearl images and their corresponding spatiogram representations. The first problem to solve then is the visualization of the multidimensional spatiogram data. Clearly, this is a non-trivial task and, to our knowledge, it has not been addressed before.

Here, we propose a spatiogram visualization scheme which involves three types of plots, the S-plot triplet (S1, S2, S3):

S1 connected centers of bins, $\boldsymbol{\mu}_b = (\bar{x}_b\ \bar{y}_b)$;

S2 $\boldsymbol{\mu}_b$–positioned counts of bins (the radii of the circles are proportional to c); and

S3 $\boldsymbol{\mu}_b$–positioned variances of bins (the lengths of x- and y-error bars are $\pm q_b(x,x)$ and $\pm q_b(y,y)$, respectively).[5]

For examples of the S-plots, we refer to Figure 6.3 (constructed example) and Figure 6.4 (actual data). In all S-plots the color identifies the bin, with a range from dark blue for $b = 1$ (the darkest pixel values) to dark red for $b = B$ (the brightest pixel values).

In Figure 6.3, two ball-shaped objects are constructed such that their histogram bins are exactly the same but the pixels from the same bins are positioned differently within the area of the object. Thus, the two objects are clearly different in their appearance; and yet they cannot be distinguished from the histograms.[6] Let us then turn to the spatiogram representations of the considered objects, which look obviously different, and see what we can infer about the objects from their S-plots.

By observing the two S1-plots, we see that the three centroids (spatial means) in the top plot are nicely grouped around the center of the plot (the center of the object area), while in the bottom S1-plot the three bin centroids are further apart. The S2-plots, in general, can be seen as a combination of the S1-plots and the histograms (hence no need for the histogram plots): the circles are centered on the bin centroids and they expand proportionally to the bin counts, a larger circle indicates a bin with more pixels. From the S2-plots in Figure 6.3, we read that the two objects have the same bin counts but differently arranged in space. Thus, the overall color tone (brightness) of the objects is the same but, for example, the brightest area in the bottom object is on average located higher than in the top object. Finally, from the S3-plots of the two constructed objects, we observe that the x- and y-coordinates of the brightest bin ($b = 1$, represented with red color) have very small variations which suggests that they are highly concentrated in space (indeed, these are the blob-like regions in the two objects). Moreover, if we would zoom-in on the red bars in the S3-plots, we would see that the red y-bar from the bottom plot is longer than its crossing x-bar suggesting a larger spread of the bin $b = 1$ pixels along the y- than along the x-axis. This again is in line with the fact that the brightest region of the bottom object forms an ellipse, unlike the brightest region of the top object which appears circular in shape and for which the corresponding x- and y-bars from the S3-plot are approximately the same in length. A similar kind of reasoning applies to the green ($b = 2$) and blue ($b = 3$) bar-pairs from the S3-plots.

Next, we move from the constructed image objects from Figure 6.3 to the images of painted objects depicted in Figure 6.4. We refer to the depicted example beads or pearls as I1, I2, I3 and I4, from left to right respectively. The spatial arrangement of

[5]Note that the definition of a spatiogram contains the covariance $\boldsymbol{\Sigma} = \begin{pmatrix} q(x,x) & q(x,y) \\ q(y,x) & q(y,y) \end{pmatrix}$. In order to simplify the presentation, we confine our analysis to the variance components $q(x,x)$ and $q(y,y)$.

[6]Remember a similar example of the real painted pearl images from Figure 6.2: the two obviously non-identical pearls which have exactly the same histograms.

pixel values in the objects now becomes notably more complex, reflected also in the S-plots.[7] Nevertheless, the interpretation of the S-plots is still pretty straightforward. For example, the S1-plots in Figure 6.4 suggest that the centroids of brighter bins are quite nicely grouped for I1 and I2, in contrast to the I3 and I4 whose S1-plots suggest larger spread of the bright pixels across the image area. On the other hand, we note that the centroids of the darkest and those of the brightest pixels are positioned very close to each other for I3 and I4, they are more far apart for I1, and they are most distant for I2. The S2-plots indicate that the darker pixels prevail in all images but there is still a noticeable portion of bright pixels in I2, somewhat less in I1 and I3, and only very few in I4. Lastly, we note from the S3-plots that pixels of the brighter bins are rather scattered across the image area of I3 and especially of I4, which is not the case with I1 and I2. In terms of the spread of their darker pixels, the four objects seem more similar to each other.

6.3.2 Exploratory visual inspection

Once we have defined the three S-plots, we can perform an exploratory visual inspection of the spatiogram data for a larger set of pearls. The goal is to determine if and which properties of the spatiograms are most informative of the pearl characteristics we are interested in; for example, symmetry of the pearl area and smoothness of its surface. For this purpose, we selected a total of about 100 painted or photographed pearl images and arranged them in several groups according to their visual appearance (in terms of size, material, apparent surface smoothness, size-shape-position of the highlight area). Example pearls from three such groups are shown in Figure 6.5.[8] The three groups differ in two main aspects:

1. the intensity range (“color palette”) - the difference is smaller between the top and the bottom group (overall, few very dark pixels) and larger between either of these and the mid group (many more dark pixels); and

2. the highlight area - the difference is smaller between the top and the mid group (the shape and size of the highlights are similar; only the “edges” are sharper for the mid group) and larger between either of these and the bottom group (the shape of the highlights is different).

While inspecting spatiogram plots of different groups of visually similar pearls, we made three main observations:

1. spatiograms within one group appear more similar than spatiograms between different groups of pearls;

[7] Obviously, the content of the S-plots is determined by the corresponding image data. Note, however, that the detailed appearance of the plots is also affected by the total number of bins B.

[8] Remember that our analysis is done in the grayscale domain. Therefore, we make all observations on the grayscale images.

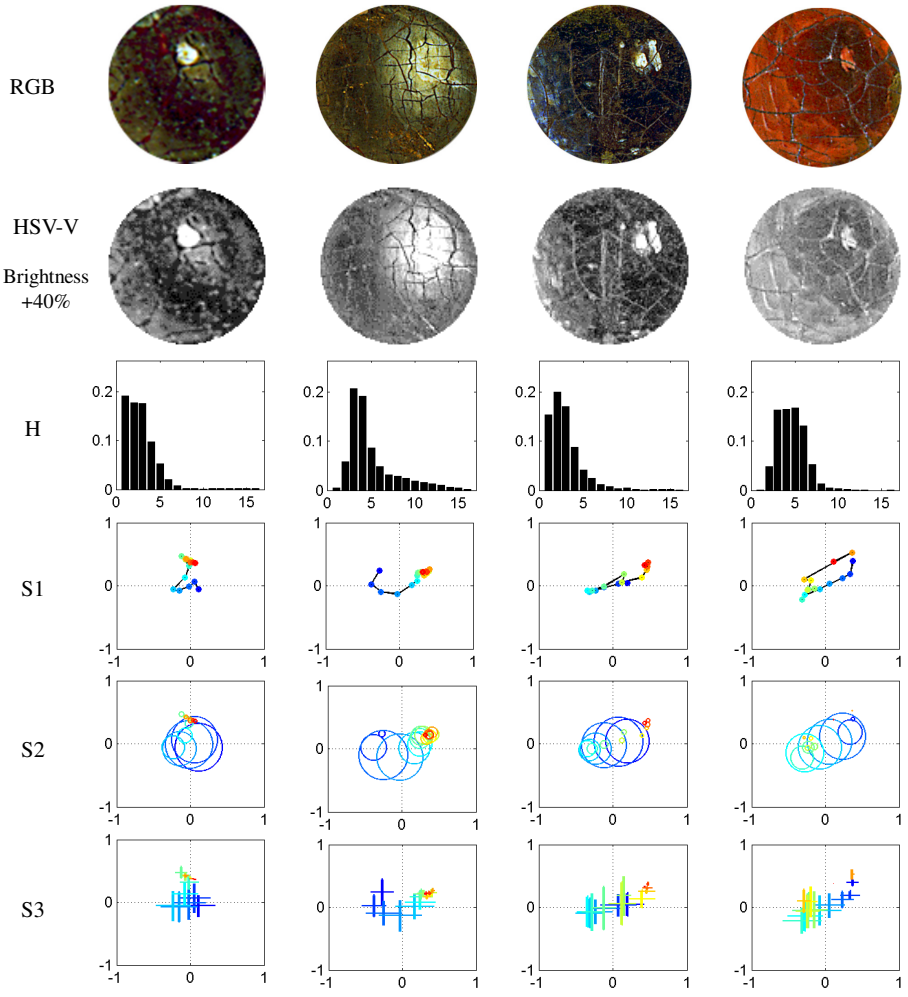


Figure 6.4: Histograms and spatiograms for example beads (columns 1,3 and 4) and a pearl (column 2) from the Ghent Altarpiece. Note the difference between specular (mirror-like) reflection of the beads (small highlight area with rather sharp edges) and the diffuse reflection of the pearl (larger highlight area with soft, blurry edges). Top to bottom: original RGB image, grayscale (HSV-V) image of the object (for the purpose of improved visualization, the brightness of the grayscale images in the figure is increased by 40 percent), histogram, spatiogram S1-, S2- and S3-plot (B=16). The color in the S-plots represents the bins, with a range from dark blue for $b = 1$ (the darkest pixel values) to dark red for $b = B$ (the brightest pixel values).

2. spatiograms between different groups are more dissimilar between groups with different highlights and similar "color palette" than between groups with different "color palette" but similar highlights; and

3. the differences (or similarities) between spatiograms are visually most obvious and easiest to interpret in S1-plots (in particular, see the curved shape of the connected centers of bins and the distribution of the specific bins along these connected lines, as well as the lengths of the connecting line segments, *i.e.*, the distances between adjacent bins).

6.3.3 Spatiogram matching based on bin-similarity

So far, we have observed empirically that the S-plots (most obviously the S1-plots) of the visually similar pearls are qualitatively similar. In order to further inform these observations, we develop an algorithm for matching the spatiogram of a given pearl image (original image/spatiogram) to that of a reference pearl (reference image/spatiogram). The goal of spatiogram matching is to transform the spatial arrangement of pixels from the original image (shuffle the pixel positions within the image area) in such a way that the spatiogram of the transformed image matches the spatiogram of the reference image, according to a given criterion. We refer to the process as the *indirect spatiogram matching* and to the resulting image/spatiogram as the *matched image/spatiogram*.

The matching can be done using a kind of Markov Chain Monte Carlo (MCMC) sampler. This means: at each step, choose a pair of pixels randomly and swap their values if the change contributes to the increased spatiogram similarity, or otherwise accept the change with a certain probability. When all the sites are visited in this way, one iteration is completed and the number of iterations depends on the desired stopping criterion. In practice, we apply first histogram matching and then we apply the MCMC sampler sequentially bin-by-bin. We define a stopping criterion in terms of the required bin similarity. In particular, we use the symmetrized Kullback-Leibler (SKL) divergence between two model Gaussians with given means and covariances, $\mathcal{N}_{1b}(\boldsymbol{\mu}_{1b}, \boldsymbol{\Sigma}_{1b})$ and $\mathcal{N}_{2b}(\boldsymbol{\mu}_{2b}, \boldsymbol{\Sigma}_{2b})$. Note here that the means and covariances of the bins are spatial and not intensity based. The two bins are considered similar enough when the SKL divergence between their model Gaussians drops below a predefined threshold. Finally, we revert to the original histogram, by applying the reverse of the initial histogram matching operation. This step preserves the original "color palette" (intensity range) chosen by the artist (and expressed in the image histogram).

Figure 6.6 shows an example of matching the spatiograms of two pearl images. The two input images into the process, the image to be matched (original image) and the reference image, are presented in the second and the first row, respectively. Note that the two images are obviously different in appearance; the same with their histograms and spatiograms. After the matching process, however, the two images

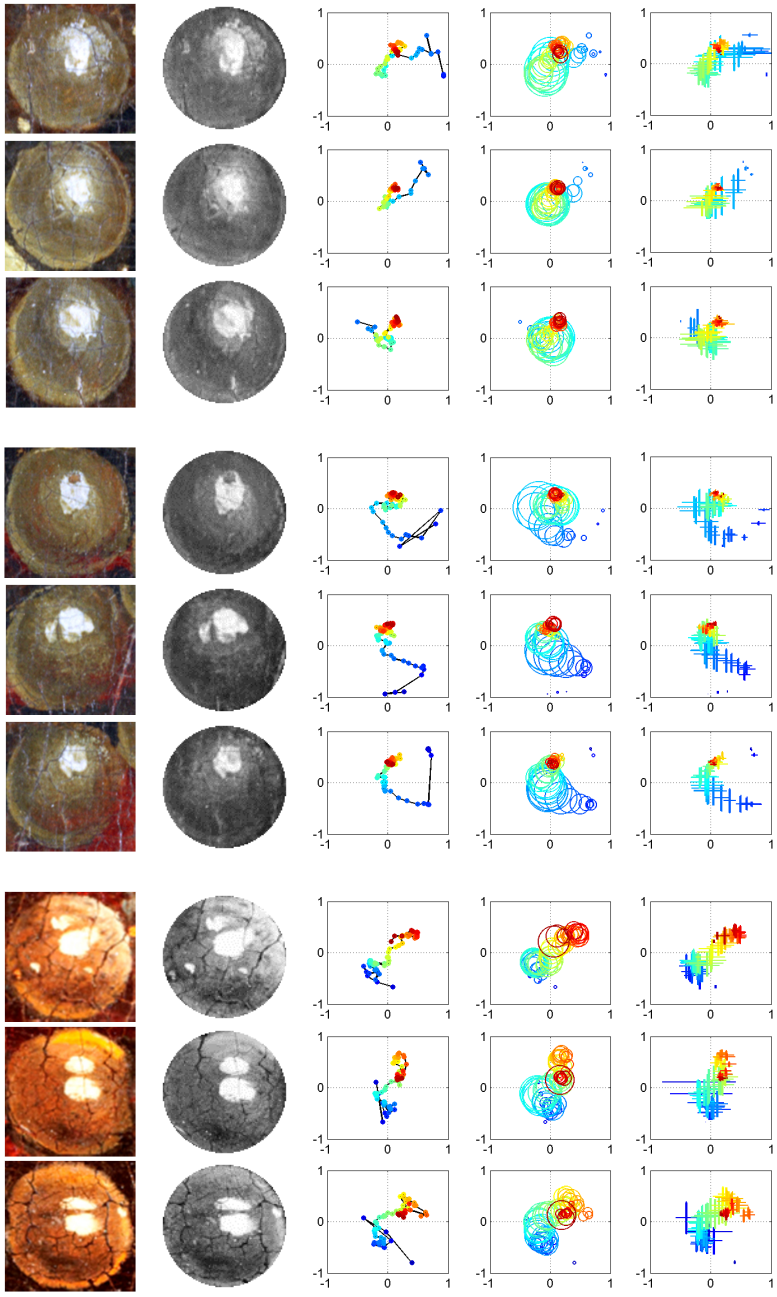


Figure 6.5: Example pearls used in the exploratory visual inspection. From top to bottom: three groups of visually similar pearls. From left to right: original RGB patch, registered HSV-V image of the pearl, spatiogram S1-, S2- and S3-plot (B=64).

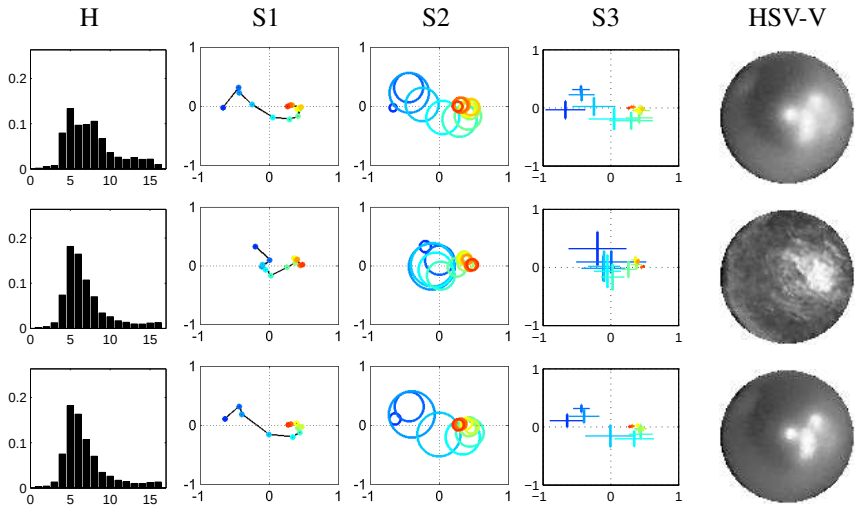


Figure 6.6: Spatiogram matching (B=16). Top to Bottom: the reference image, the test image before matching (original image), and the matched image (the result of matching the test image to the reference). Left to right: histogram, S1, S2, S3, and registered HSV-V pearl.

become highly similar for their appearance. By comparing the grayscale image from the bottom (matched image) to that from the top row (reference), we note hardly any differences. While the histogram of the original image remained unchanged with the process of spatiogram matching, the S-plots have notably changed - they are now highly similar to the S-plots of the reference. Here again, the same as in Section 6.3.2, S1-plots seem to most clearly reflect the similarities between the pearls. Therefore, and given the early stage of this research overall, we decide to retain only the S1-plots for our further analysis.

In the final step of our investigation, described next , we formulate a set of four numerical measures which quantify the observed relationships between the appearance of pearls and their image spatiograms (that is, their S1-plots).

6.3.4 New spatiogram-based measures of object appearance

By visually analyzing the S1-plots of many painted pearls as well as of their altered versions obtained by spatiogram matching to a range of arbitrarily selected pearls, we found four new measures that characterize some important features of the appearance of (painted) pearls in 2D images. The proposed measures are derived from the spatial centers of intensity bins, $\mu_b = (\bar{x}_b\ \bar{y}_b)$, and their distances, D_b. Especially, we define the distance between the centers of the adjacent bins as the Euclidean distance:

$$D_b{}^2 = (\bar{x}_{b+1} - \bar{x}_b)^2 + (\bar{y}_{b+1} - \bar{y}_b)^2, \tag{6.8}$$

where $b = 1, ..., B-1$. Then, the four new measures are defined as follows:

$$\mathrm{DistMean} = \frac{1}{B-1}\sum_{b=1}^{B-1} D_b, \tag{6.9}$$

$$\mathrm{DistStd} = \frac{1}{B-1}\sum_{b=1}^{B-1} (D_b - \mathrm{DistMean})^2, \tag{6.10}$$

$$\mathrm{RangeX} = \max_b(\bar{x}_b) - \min_b(\bar{x}_b), \tag{6.11}$$

$$\mathrm{RangeY} = \max_b(\bar{y}_b) - \min_b(\bar{y}_b). \tag{6.12}$$

For a discussion of the physical meaning of the proposed measures, we create simplified model images as shown in Figure 6.7. Note that these synthetic images are meant to demonstrate the properties of the proposed measures and, while motivated by the basic facts about pearl objects, they do not necessarily depict true physical properties of the real pearls. We limit the number of bins in the images to three, inspired by the three main areas of a pearl [Farn, 1986]: the often observed mirror image of the light source (b_3), the main surface of the pearl (b_2), and the glowing sheen usually present against the outline of the pearl (b_1). For a visual description of the main parts of a pearl image, see also the drawing in Figure 6.18.

The $\mathrm{DistMean}$ and $\mathrm{DistStd}$ measures are defined in Eq. (6.10) and Eq. (6.11) as, respectively, the mean and the variance of the distances between adjacent bin centroids. $\mathrm{DistMean}$ reflects the symmetry of the pearl area. Given the definition of the spatiogram centroid (geometric center of the pixels from a given bin) and the geometry of an imaged pearl (approximately circular in shape), the minimum value of $\mathrm{DistMean} = 0$ corresponds to the case where all centroids are positioned at the center of the pearl area (the point (0,0) in the S1-plot). In that case, for example the pearl in Figure 6.7 (a), the centroid overlaps with the (approximate) center of symmetry of the pearl area, that is, the center of symmetry is the center of the pearl area. Likely, in the specific case where $\mathrm{DistMean} = 0$, also the highlight area would be positioned in the center of the pearl.[9] The position of the highlight is of interest because it holds clues about the position of the light source in the scene:[10] the closer the highlight to the center of the pearl, potentially the less sharp angle between the light source and the pearl surface. If we now consider the other three examples in Figure 6.7, $\mathrm{DistMean}$ is the smallest for (b) in which two out of three centroids overlap and it is the largest for (d)

[9] Note though that this is not conditioned by the value of $\mathrm{DistMean}$. For example, we could think of an image in which the geometric center of the brightest pixels is in the image center but they are scattered "randomly" throughout the image. Related discussion concerning the modelling of the bin data coordinates can be found in Section 6.2.2. Evidently, for a general application with no assumptions (prior knowledge) about the properties of the object, a more confident inference about the grouping of the pixels would require the S3 information included in the analysis. For our purposes, we assume that the highlight area is always painted the brightest and at a single location. At this moment, the S2- and S3-based measures are left for future research.

[10] For art history studies, it is of interest to observe, for example, if the lighting of the scene is captured consistently throughout the panel, or if there are some patterns in how the optical effects are rendered.

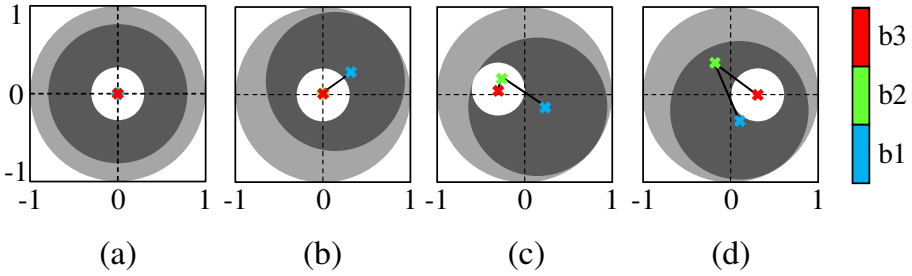


Figure 6.7: The images to illustrate the properties of the proposed measures DistMean, DistStd, RangeX, and RangeY defined in Eq. (6.10) – Eq. (6.12).

where the centroids are further from each other. At the same time, we notice that the symmetry of the objects has changed. Compared to (a), there is now less symmetry in (b), even less in (c) and clearly much less in (d).

The next DistStd measure, relates to the uniformity of "distances" between different bin areas, or the impression of the surface uniformity.[11] For example, observe the objects in Figure 6.7 (c) and (d). In object (d) (smaller DistStd), there are two types of transitions, one between $b3$- and $b1$-area and the other between $b1$- and $b2$-area. In object (c) (larger DistStd), however, there is an additional seemingly "disturbing" transition between $b1$- and $b2$-area which affects the appearance of surface uniformity. With this in mind, if we compare DistStd of object (d) to DistStd of (a), it suggests larger uniformity of (a). Indeed, despite the same transitions in the two objects, the areas of different bins are distributed more uniformly in (a) than in (d).

Lastly, the RangeX and RangeY measures describe the range of bin centers in x- and y-direction, respectively. These two measures tell us about the (dominant) orientation of the asymmetry in the pearl, if any. For example, if a blurry highlight representing a mirror image of the light source would be further away from the central vertical axis of the pearl (sharper angle between the light source and the pearl surface), we would expect larger RangeX compared to the case where this highlight is more centered on the pearl area (see objects (c) and (d) versus object (b) in Figure 6.7).

In this work, the four proposed measures are used as separate descriptors of different appearance attributes. In the future, especially after the set of measures has been extended to depict also other attributes of appearance (*e.g.* those described by the S2- and S3-information), it may be of interest to develop a method for appearance-based classification of pearls (or other objects), using these attribute measures as classifier features.

[11] The research of perception and psychophysics suggests that surface uniformity is an important aspect of object-based attentional selection [Watson and Kramer, 1999]. [Chen, 2012] write "All else being equal, a 'good' object is one that has surface uniformity and closed boundaries."

6.4 Experimental results

Despite the now readily available technology for artwork digitization, this process is overall still in its infancy. In the summary [Farnand et al., 2013] and final report [Frey and Farnand, 2011] of a three-year project on "Current Practices in Fine Art Reproduction" the authors motivate their work by noting that "To create reproductions of their artwork, cultural heritage institutions employ a range of technology and a variety of workflows." The same is true for the images used in our study. The results presented in this section are obtained for digital photographs of the paintings collected from different sources and acquired under different, often non-controlled conditions, either by the professional photographers or by amateurs. Another important consequence of the early stage (and the rather slow progress) of the digitization efforts in art is the fact that a vast majority of the current digital records are incomplete, comprising only selected pieces from the collections. For this reason, some of the experiments in our study could only be performed for a limited number of painted pearls or beads. Finally, we remark that much of the digital art materials are proprietary to the institutions and/or private owners (also in our study) which often creates difficulties in gaining access to the data, even for research purposes.[12]

6.4.1 Automated image processing system

We implement a fully automated system for our experiments, as illustrated in Figure 6.8. The spatiogram-based analysis is preceded by a range of preprocessing steps. For pearl detection, we use the Hough transform [Duda and Hart, 1972], iteratively for a set of radii of interest. To reject false detections, for smaller pearls we also use a set of features that characterize painted pearls (the angle of reflection, smoothness and mean gray value). The images of extracted pearls are transformed to HSV color space and further on only the Value (V) channel is used (pixel intensity range 0-255). To eliminate concerns about the influence of cracks on the proposed pearl analysis, we performed crack detection [Cornelis et al., 2013] followed by crack inpainting [Ružić et al., 2011]; however our results (not shown here) suggest that spatiograms of pearl images are robust in handling cracks.[13] The last preprocessing step is the registration of the pearl images to the "reference" pearl (one arbitrarily selected pearl from the

[12] Most recently, shortly after our study was conducted, an open access digital record of the *Ghent Altarpiece* appeared. The images were acquired under controlled conditions, in a well-defined and fully documented process, and the full record is publicly available through the website of "Closer to Van Eyck: Rediscovering the Ghent Altarpiece", `http://closertovaneyck.kikirpa.be`. The database was created in the scope of the *Lasting Support*, an interdisciplinary research project (April 2010 - June 2011) to assess the structural condition of the Van Eycks' XVth century master piece from Saint Bavo Cathedral in Ghent, Belgium. Importantly, however, the database does not include a scan of the *Just Judges* panel.

[13] We remark, though, that the effects of cracks on the perceived appearance of pearls have not been investigated in this study. It is possible that human observers are more sensitive to the cracks than the mathematical methods explored here. Should that be the case, the current methods for image analysis revised accordingly.

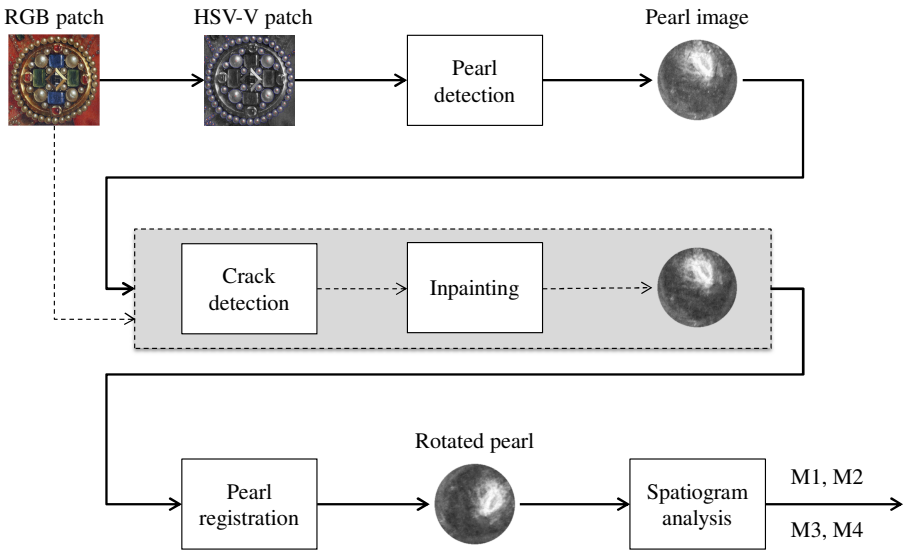


Figure 6.8: Block scheme of the system for automated pearl image analysis.

considered set). The registered pearls are then subjected to spatiogram analysis using the numerical measures described in Sections 6.2.2 and 6.3.4.

6.4.2 Pearls and beads in the *Ghent Altarpiece*

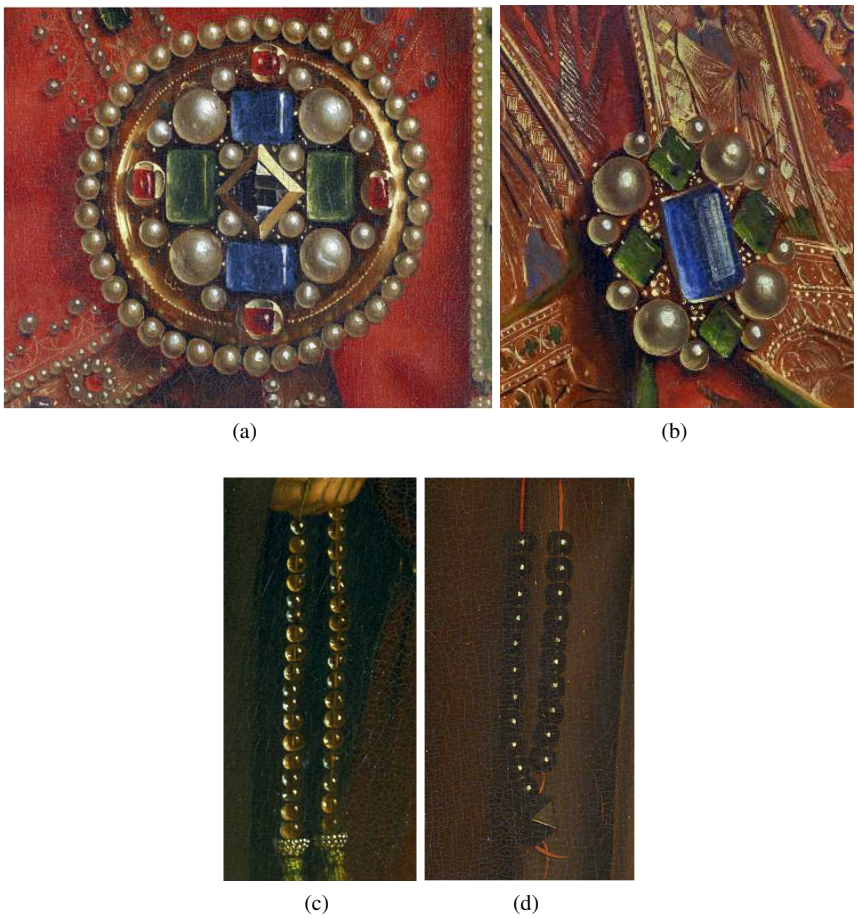
The *Ghent Altarpiece* depicts a great number of pearls, beads and other spherical ornamental objects. In the following analysis, we consider three panels from this polyptych: *God the Father* (Figure 6.9), *Singing Angels* (Figure 6.10) and *Holy Hermits* (Figure 6.11). Figure 6.12 shows close-ups of the specific selected details with pearls and beads from these three panels. We assess the similarity of different objects using the spatiogram similarity measure ρ from Eq. (6.7). The results obtained for $B = 128$ bins are summarized in the four bar charts shown in Figure 6.13. Each plot shows the percentage of all paired comparisons of pearls, or beads, for which the similarity ρ falls in the given interval, $\rho \in [0.1(k-1), 0.1k]$, $k = 1, 2, \ldots, 10$. The title of each plot names the specific image detail(s) under analysis.

Several observations can be made based on the charts in Figure 6.13. First, we analyze the top two charts which measure the similarity between pearls of a single painted ornament. The top left chart shows the binned ρ values computed for all paired combinations of a total of 51 pearls lying in the frame of the broach in Figure 6.12 (a). The top right chart shows the results of the same comparisons but between the pairs of the 70 pearls on the coat, the smaller pearls located outside the broach area. We refer to the 51 pearls in the broach frame together as *Object 1* and to the 70 pearls outside the broach as *Object 2*. Note that the pearls of Object 1 are larger than those of Object

Figure 6.9: Pearls in the *Ghent Altarpiece*: the brooch in the *God the Father* panel.

Figure 6.10: Pearls in the *Ghent Altarpiece*: the brooch in the *Singing Angels* panel.

Figure 6.11: Beads in the *Ghent Altarpiece*: the rosaries in the *Holy Hermits* panel.



(a) (b)

(c) (d)

Figure 6.12: Pearls and beads in the *Ghent Altarpiece*, details from: (a) *God the Father*, (b) *Singing Angels*, (c) and (d) *Holy Hermits*. For images of the corresponding whole panels, see Figures 6.9, 6.10, and 6.11.

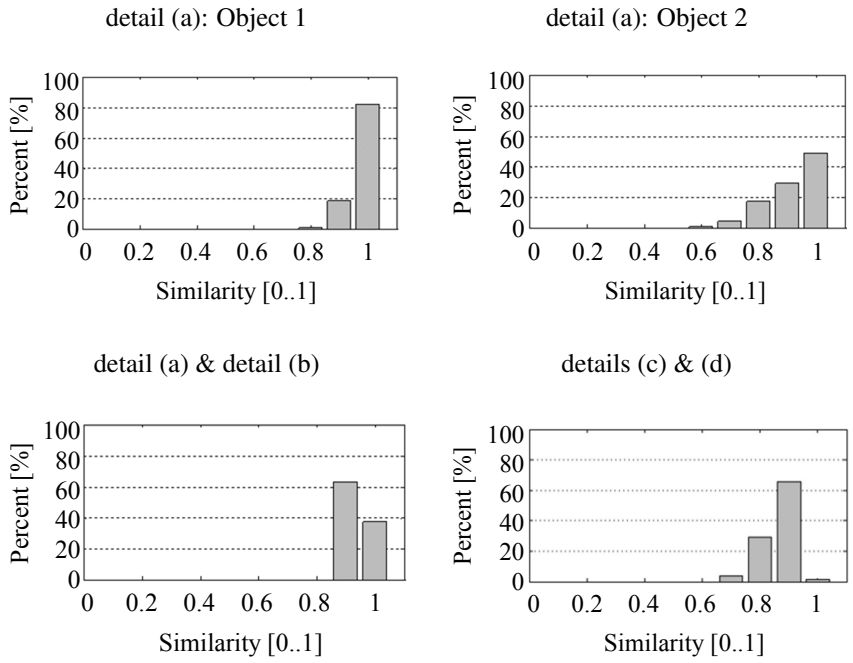


Figure 6.13: Similarity of pearls and beads from the details in Figure 6.12 (see text).

2. The described comparisons between pairs of pearls which all belong to the same object are referred to as *intra-object* comparisons. If we now compare the two top plots in Figure 6.13, we notice that the intra-object similarities are notably larger for Object 1 than for Object 2. In Object 1, approximately 80% of the pearls are similar to each other with as much as $\rho > 0.9$, contrasted to $\approx 50\%$ among the pearls from Object 2. This suggests that the artist painted larger pearls (those from more salient objects, such as Object 1) with more care and more attention to detail than the smaller pearls (which probably attract less direct visual attention, such as Object 2). This is, of course, not unexpected but it would be of much interest to compare in this respect different pieces of artwork by the same artist (for example, to answer the questions such as "Does the artist always give the same attention to the objects of a given size?" or "Does the artist assign special attention to a specific object?"), and especially to compare different artists.

In light of these interesting challenges, we now compute the similarity index ρ for all pairs of pearls from different panels (objects) and refer to these comparisons as the *inter-object* comparisons. We measure the similarity between the largest 4 pearls from the detail in Figure 6.12 (a) to the four pearls from detail in Figure 6.12 (b). These are pearls of the similar size but from two different panels, *God the Father* and *Singing Angels*, respectively. The results shown in the bottom left plot of Figure 6.13 indicate

the inter-object similarity of $\rho \geq 0.8$. By comparing this to the intra-object pearl similarities from the top two plots, we reason as follows. In the *Ghent Altarpiece*, the larger pearls which belong to a visually salient object are highly similar; for the less salient objects and the smaller pearls, the extent of similarity slightly drops. Quite impressively, the visually salient pearls of similar size are painted fairly similarly even across different panels.

Lastly, we analyze the bottom right plot in Figure 6.13. The bars describe the ρ values for inter-object similarity between the 21 glass beads from the detail in 6.12 (c) and the 9 wooden beads from 6.12 (d). We notice that only very few of the glass beads were similar to the wooden beads ($\rho > 0.9$ for less than 5% of the comparisons). The relative comparison of this versus the other three bar charts in Figure 6.13 suggests that different materials are painted notably differently.

6.4.3 Pearls from different artworks: How do they differ?

With this experiment, we aim to evaluate the numerical measures from Sections 6.2.2 and 6.3.4 for their ability to inform about some specific attributes of the visual appearance of a painted pearl. As elaborated in the beginning of this chapter, such a description of the painted objects would be a very useful argument for art historical analyses; for example, measurements like this may serve as an important indicator of the ability of an artist to observe nature.

For this analysis, we consider one representative pearl from each of the following art works: (Pearl 1) from *God the Father* in the *Ghent Altarpiece* painted originally by Hubert and Jan van Eyck in the XVth century, (Pearl 2) a copy painted by Charlotte Caspers (2010) of the *Angels Playing Music* from the *Ghent Altarpiece*, (Pearl 3) a copy painted by Jef Van der Veken (1945) to replace the stolen panel *The Just Judges* from the *Ghent Altarpiece*, and (Pearl 4) one of the masterpieces of Northern Renaissance art *Maria Maddalena Baroncelli* painted by Hans Memling (1470). In addition, we consider one photograph of a real (rather than painted) pearl taken from the test set of ten real pearl photos (Pearl 5).[14] The five pearls, shown in Figure 6.14, are selected according to two criteria: (1) the pearl image size is within a given range (100 ± 30 pixels, the most common range for our data set), and (2) the pearl image spatiogram is representative of its class (the painted ornament, the panel, or the set of photographs, from which the pearl is taken).

Figure 6.14 depicts the test pearls in color (original image data) as well as in grayscale (after the preprocessing described in Section 6.4.1). Also shown are the histograms and the spatiogram plots for the grayscale images. Note here again that, as established earlier in this chapter, the histograms of the pearls are not directly informative of the visual appearance of the objects. They only describe the distribution of the color tones ("color palette") in the painted object.

[14] Note that the type of light source in Pearl 5 is different from that in Pearls 1-4.

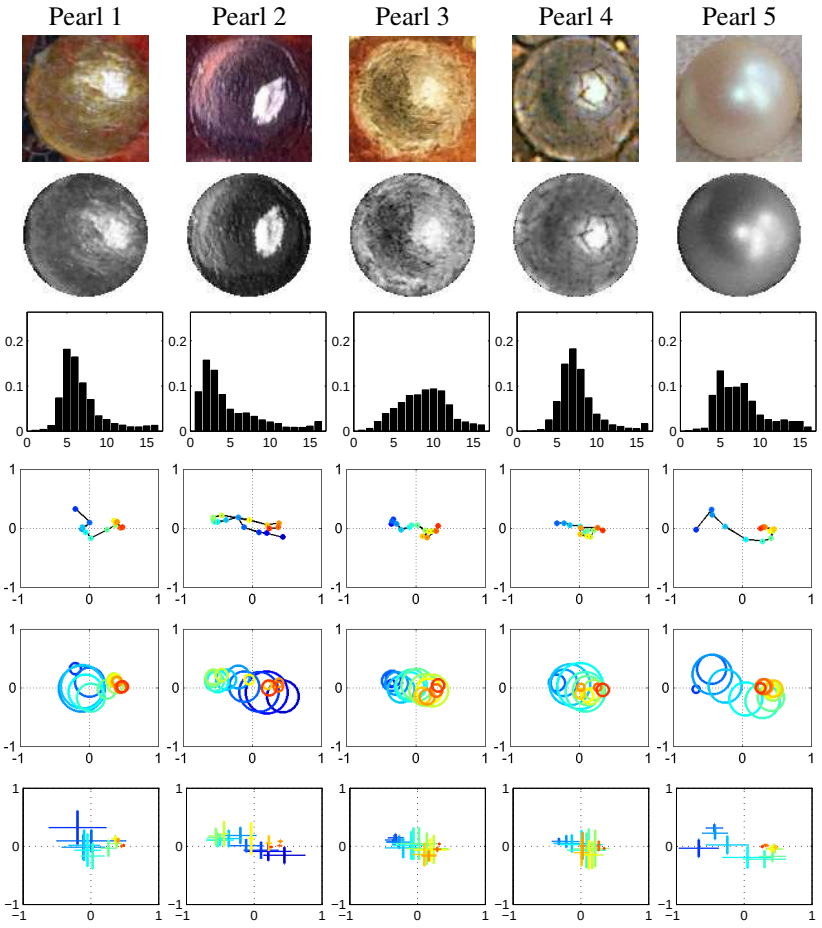


Figure 6.14: Using spatiograms to characterize the appearance of pearls in the images. Top to bottom: original RGB patch, registered HSV-V pearl, histogram, S1-plot, S2-plot, and S3-plot (B=16).

Table 6.1: Spatiogram similarity measure ρ in comparing pearls from Figure 6.14

	Pearl 1	Pearl 2	Pearl 3	Pearl 4	Pearl 5
Pearl 1	1	0.726	0.851	0.906	0.904
Pearl 2	0.726	1	0.730	0.819	0.673
Pearl 3	0.851	0.730	1	0.849	0.879
Pearl 4	0.906	0.819	0.849	1	0.849
Pearl 5	0.904	0.673	0.879	0.849	1

First, we observe the values of the existing spatiogram similarity measure ρ defined by Eq. (6.7) computed for all pairs of pearls from Figure 6.14, shown in Table 6.1. For example, if we compare Pearls 1-4 to Pearl 5 (see the last column or the last row of Table 6.1), the value of ρ is highest for Pearl 1 ($\rho_{1,5} = 0.904$), it is somewhat smaller for Pearl 3 ($\rho_{3,5} = 0.879$) and Pearl 4 ($\rho_{4,5} = 0.849$), and it is smallest for Pearl 2 ($\rho_{2,5} = 0.673$). This suggests that Pearls 1, 3, 4 and 5 are quite similar while Pearls 2 and 5 are rather different. However, it tells us nothing about *how* the pearls differ. Are their color tones (palette) different? Or is it because one pearl is smooth and the other appears very bumpy? Or maybe even their highlights are different, *e.g.*, one pearl has a tiny highlight in its center and the other pearl's highlight spreads along half of the pearl boundary? As evidenced by the analysis presented in Section 6.4.2, the kind of information provided by the ρ measure can be enough to support the argument about how the artist divided his or her attention between the objects of interest (very similar objects received more attention than the dissimilar ones). However, if we now want to determine the details of the differences between the observed pearls (think, for example, of when we want to compare two different artists and their painterly executions), the ρ measure is no longer able to serve the purpose. That is, we can use ρ to measure the extent of the differences (smaller or larger) but we can not explain them; the ρ measure is does not provide information about the details of the captured differences.

Then, we resort to the novel set of measures $\mathrm{DistMean}$, $\mathrm{DistStd}$, RangeX, and RangeY proposed in Section 6.3.4. These results are shown in Table 6.2. The values of $\mathrm{DistMean}$ suggest that for Pearl 3 and Pearl 4 (the smallest $\mathrm{DistMean}$) the symmetry of appearance is higher than for Pearl 5 and Pearl 1. Also according to $\mathrm{DistMean}$, the pearls are ordered the same for the position of their highlight. In Pearl 3 and Pearl 4, the highlight is painted closer to the center while in Pearl 5 and Pearl 1 it moves closer to the boundary of the object. Similar observations can be deduced from the RangeX values which quantify the spread of the centroids in the x-direction. The RangeX suggests the narrowest x-range for the centroids for Pearl 3 and Pearl 4 and the widest for Pearl5 and Pearl 1. Note that RangeY values are not very different between the five pearls. This is expected since the pearls have been registered during the preprocessing stage such that their highlights are approximately aligned, lying

Table 6.2: New spatiogram measures in characterizing pearls

	Pearl 1	Pearl 2	Pearl 3	Pearl 4	Pearl 5
DistMean	0.231	0.185	0.122	0.146	0.210
DistStd	0.063	0.016	0.010	0.015	0.047
RangeX	1.358	0.998	0.692	0.777	1.347
RangeY	0.523	0.364	0.332	0.395	0.544

Table 6.3: New spatiogram measures in identifying artists

	Van Eyck	Van der Veken
DistMean	0.210± 0.029	0.136± 0.013
DistStd	0.055± 0.019	0.009± 0.002
RangeX	0.767± 0.229	0.733± 0.028
RangeY	0.899± 0.210	0.457± 0.139

along the x-axis. Conveniently, these suggestions all seem to align reasonably well with what we can observe on the images with the naked eye.

Finally, and not unexpectedly, we can see that the conclusions concerning the similarity of the pearls drawn from Table 6.2 (the proposed measures) do not always agree with those drawn from Table 6.1 (ρ measure). This is further discussed in Section 6.4.5, after we analyze the data from our experiment with humans.

6.4.4 Who painted the pearls?

In order to evaluate the potential of the proposed measures to discriminate between pearls of different artists, we look back in the *Ghent Altarpiece* and compare the pearls by the Van Eycks' to those by Van der Veken. In particular, we select

1. 20 Van Eycks' pearls from Object 1 in detail (a) of Figure 6.12 whose spatiograms are most similar ($\rho > 0.8$), and

2. the 4 pearls from Van der Veken's copy of the *Just Judges* panel from the *Ghent Altarpiece*; see Figure 6.15.

The mean and standard deviation of the measures DistMean, DistStd, RangeX, and RangeY for these two selections of pearls are summarized in Table 6.3. The results clearly indicate the difference between the two artists' hands, that is, between their painterly executions of pearls [Verougstraete et al., 2004].

Finally, Figure 6.16 shows the results of matching pearls of other artists to the Van Eyck's pearl. This kind of analysis can be of interest in, for example, studying the influence of the pearl characteristics on the visual impression of a painting.

Figure 6.15: Pearls in the *Ghent Altarpiece*: the brooch in the *Just Judges* panel.

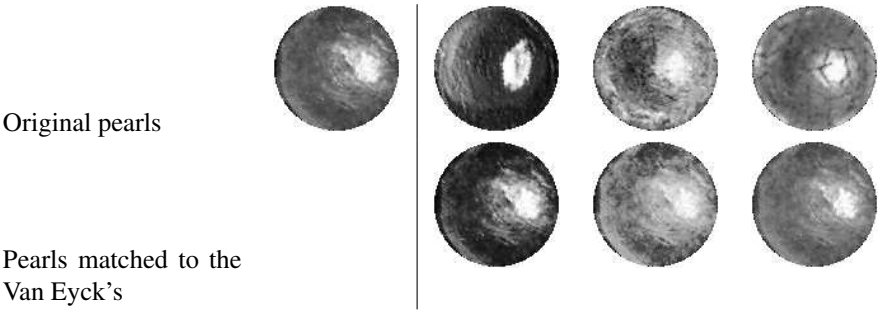


Figure 6.16: Top (left to right): Pearls of Van Eyck, Caspers, Van der Veken, and Memling. Bottom: the upper pearls after spatiogram matching to the Van Eyck's pearl.

6.4.5 Human experiments

We conducted a small experiment with humans to collect their ratings of similarity between the pearl images from Figure 6.14. An example question sheet from this experiment is shown in Figure 6.17. The participants were asked to rate the similarity of a total of 10 pairs of pearls (each of the five pearls were compared to each other) using a 6-point discrete scale (from 0 for low similarity to 5 for high similarity). The images were presented in printed form: 10 question sheets were printed on paper, then ordered in a random fashion and stitched into an "experiment book". In addition, a questionnaire sheet was included at the end of each experiment book asking the participants to indicate which features they found most relevant for judging the similarity of the image pairs; see Figure 6.18.

For the purpose of this experiment, we assume that the printing process is "ideal", *i.e.*, the potential artifacts induced by the printing process have no major effect on the visual appearance of the pearl objects in the images. Clearly, a more rigorous scientific evaluation in the future would require a strict procedure for quality control of the digital image presentation (display or printing) to be used in the human observer experiments.

The total number of participants was 41: 17 of them were participants in The 4th Image Processing for Art Investigations[15] (IP4AI) workshop in September 2011 and the other 24 participants were participants in the meeting of the research group of Ghent University Association on High-Performance Embedded Systems[16] (in Dutch, "Krachtige Ingebedde Systemen", KIS) in November 2011. Considering their (assumed) expertise and experience in artwork and art investigations, we refer to the two groups of participants as "expert" and "non-expert" subjects, respectively. The experiment sessions, with both the experts and the non-experts, were preceded by a talk about digital image analysis of pearls and thus the subjects were introduced to the

[15] `http://ip4ai.org/`
[16] `http://kis.elis.ugent.be/`

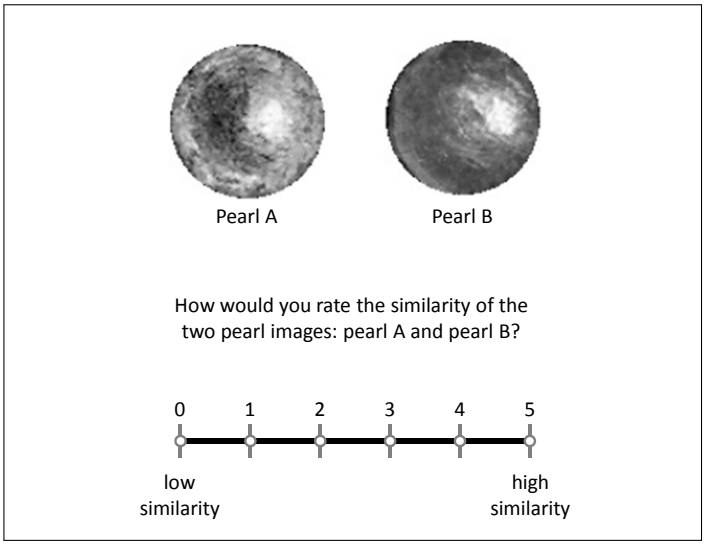


Figure 6.17: An example question sheet from the experiment with humans. The pearls were the five grayscale images from Figure 6.14 and all pairs were compared. The experiment book consisted of 10 sheets with 1 sheet per pair of pearls, ordered in a random fashion.

topic and presented with the goals of the research as well as the goals of the experiment itself.

The results of the human similarity ratings are presented in Figure 6.19, in the form of histograms, and in Figure 6.20, as boxplots. Overall, by comparing the results of the two subject groups, experts and non-experts, we notice that the level of agreement between subjects was perhaps slightly higher for the experts (less variation in the ratings for a single pair of pearls). Nevertheless, according to the median rating per pair pooled over both groups of subjects, as shown in the bottom plot of Figure 6.20, the end ranking of pairs for their degree of similarity was the same. The most similar pearls were those from the following pairs: 1-4, 1-5, 3-1, 3-4, and 5-4. We note also that the difference between the highest (most similar) and the lowest (least similar) median rating is rather small (3 in comparison to 1, on the scale of 0 to 5). This may suggest that the subjects were either very cautious about their ratings and opted to take the conservative approach, or they had difficulty defining the criteria for their judgments. Finding the most appropriate protocol for this kind of experiments is definitely an interesting question for future research.

By comparing the human and numerical ratings (the values of the similarity measure ρ from Table 6.1), we can see that the agreement between humans and the measure is quite good. The only exception is the pair of Pearl 3 and Pearl 5 which humans rated as "medium" similar while the numerical method assigned a high similarity in-

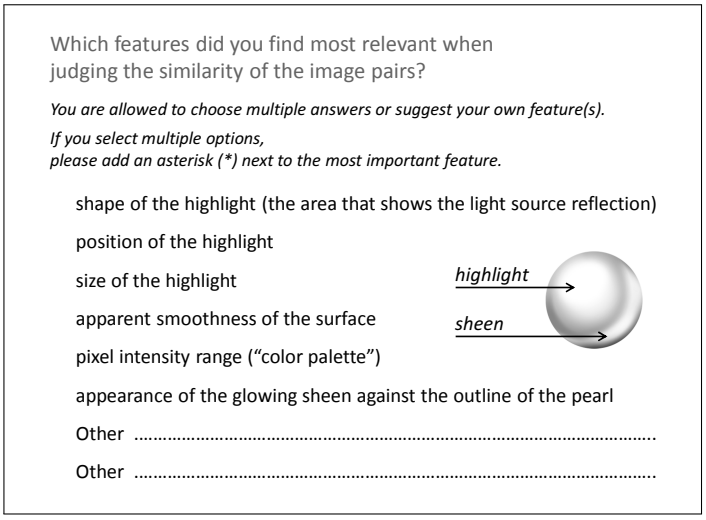


Figure 6.18: The questionnaire from the experiment with humans. After they have rated the similarity of all 10 pairs of pearls, the participants were asked to indicate which of the suggested features, if any, they thought had most influenced their similarity judgments. Suggesting a new feature, not included in the predefined list, was also allowed.

dex $\rho_{3,5} = 0.879$. While this needs to be examined with a larger data set, it is quite a promising finding. Albeit the ρ measure is unaware of the properties of the human visual system (HVS), it could prove to be a good predictor of human similarity judgments.

The last result of this chapter is presented in Figure 6.21, the bar chart of human responses to the question "Which features did you find most relevant when judging the similarity of the image pairs?" Humans seem to agree that the position of the highlight and the smoothness of the surface, which are addressed by our proposed $\mathrm{DistMean}$ and $\mathrm{DistStd}$ measures are among the most important features. Nevertheless, the list does not end with these two attributes of appearance. Some other potentially contributing factors include the shape of the highlight and also the size of the highlight. And, quite likely, the list could be extended further.

Recall the observation at the end of Section 6.4.3 that the set of novel measures versus the ρ measure may not necessarily arrive at the same conclusions about the overall similarity of pearls. The reason is exactly in these other contributing attributes that appear to be influencing the overall perception of similarity but which are not quantified by the current set of appearance measures. Therefore, it is the challenge for future work to identify other key features of (pearl) appearance and develop algorithms to quantify them.

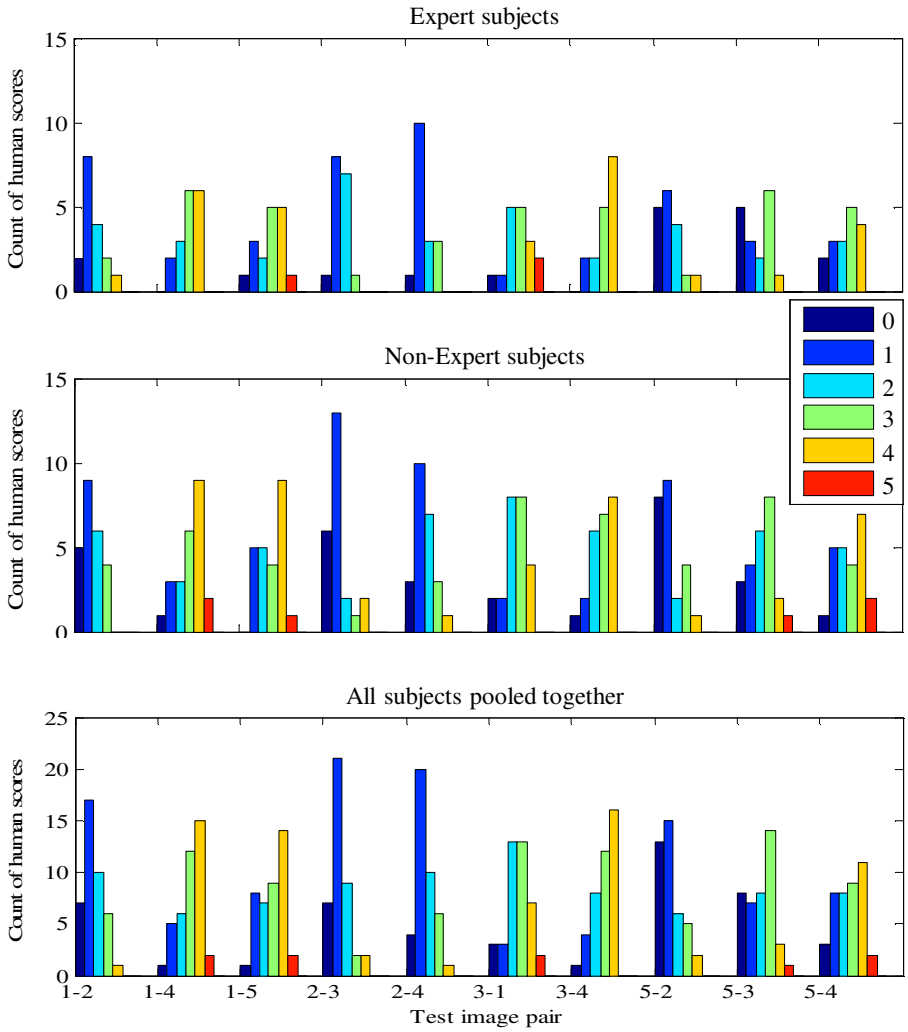


Figure 6.19: Histograms of the human similarity ratings for the five grayscale pearls from Figure 6.14. From top to bottom: ratings of the expert subject group, ratings of the non-expert subject group, ratings of the two subject groups pooled together. A higher score indicates higher similarity.

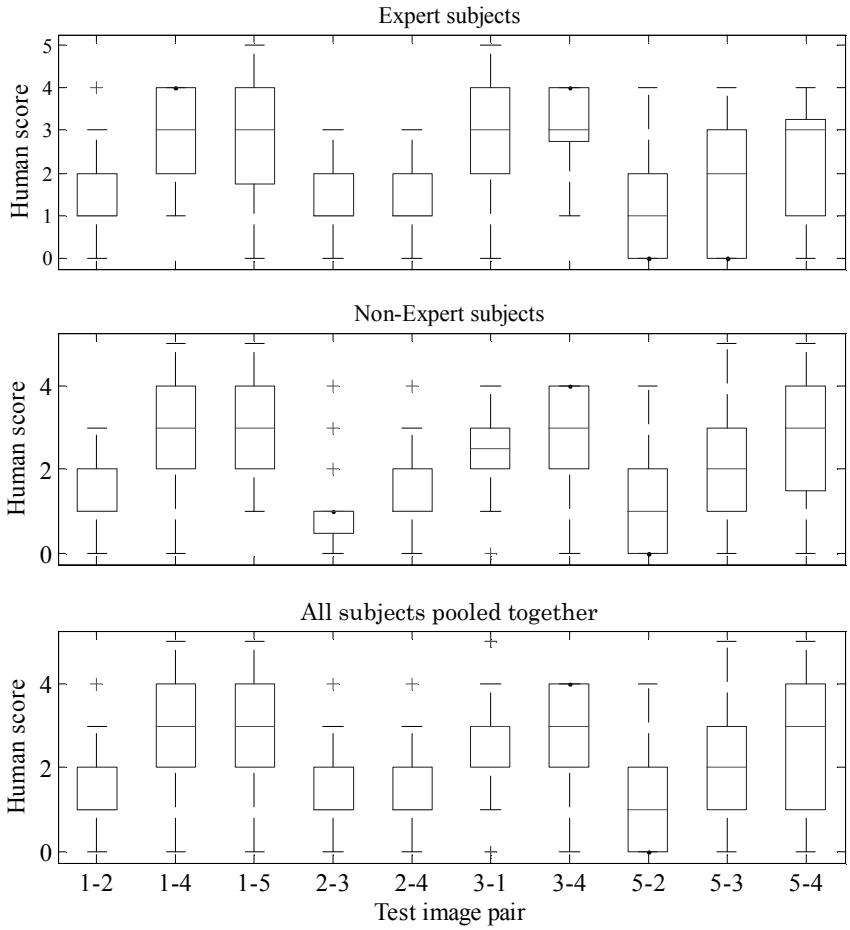


Figure 6.20: Boxplots of the human similarity ratings for the five grayscale pearls from Figure 6.14. A higher score indicates higher similarity. The central mark on each box indicates the median, the edges of the box are the 25th and 75th percentiles, and the whiskers denote the range of the data not considered outliers (which are denoted by "+" marks). From top to bottom: ratings of the expert subject group, ratings of the non-expert subject group, ratings of the two subject groups pooled together.

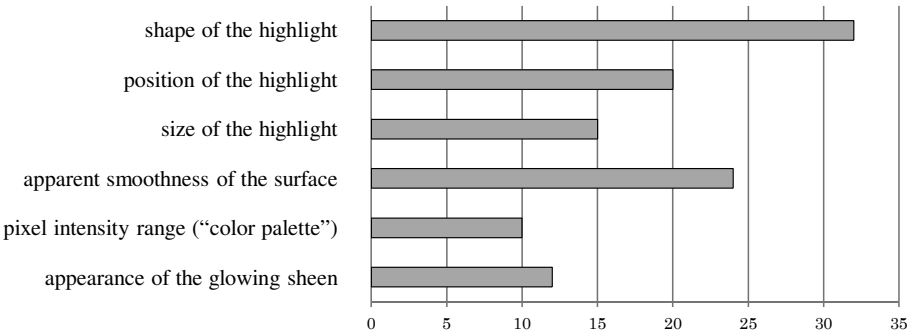


Figure 6.21: Results from the questionnaire in Figure 6.18 which asked human subjects to indicate the features which they thought had most influenced their pearl similarity judgments. They could choose as many attributes as they felt applied. The data are pooled over all the subjects, the expert and the non-expert group together.

6.5 Conclusion

The work reported in this chapter focused on developing methods for numerically quantifying attributes of appearance of pearls and pearl-like objects in 2D images. The spatiogram representation of the image data served as the framework for our analysis. It has been chosen as a means of incorporating the information about spatial distribution of image pixel intensities.

Our first contribution is the proposed method for visualizing the multidimensional spatiogram data; the problem which has not been addressed before. Next, we studied a spatiogram similarity measure suggested by the literature and found a good concordance between the measure and the human judgments of similarity between pearl images. However, when the pearls were dissimilar, the existing similarity measure was not able to provide insight into the specifics of the differences. As an example, the measure could not indicate if the pearls differed in the smoothness of their surfaces, or in the size of their highlight areas; the information which is essential for the art historical analysis. To address that problem, we developed a method for matching spatiograms of different images. This was used as a toll for our explorative analysis of the relationship between the dominant factors of the appearance of pearls (and pearl-like objects) in images and the properties of the corresponding spatiograms. Lastly, based on the observations from our explorative analysis, we proposed a set of novel spatiogram-based measures which quantify numerically the appearance of surface smoothness and several attributes regarding object symmetry. The methods have been evaluated on images of painted as well as of real pearls. Overall, the observed agreement between the new measures and visually observed image features makes the proposed approach a promising candidate for practical use in characterizing pearls in paintings. In the present work, the four proposed measures were used as separate de-

scriptors of different appearance attributes. In the future, especially after the set of measures has been extended to depict also other attributes of appearance (*e.g.* those described by the S2- and S3-information), it may be of interest to develop a method for appearance-based classification of pearls (or other objects), using these attribute measures as classifier features. Several specific directions for future research are suggested in the conclusion chapter of the thesis.

Tentative applications for the proposed techniques and their advancements include the following: (1) assisting art historians in better understanding the differences or similarities between different artists and their ways of painting pearls, (2) artist identification, and (3) forgery detection (perhaps relying more on non-appearance spatiogram features in order to better assess the differences which are imperceptible to humans).

Beyond the domain of artwork analysis, these kinds of techniques could be extended to medical image applications. In lung images, for example, the nodules are typically spherical and the degree of their uniformity is an important argument in assessing the pathology. Another example are dermatology images where, on the one hand, it is of interest to assess the fidelity of images, and on the other hand, to characterize specific attributes of the appearance of a skin lesion.

The contributions reported in this chapter have resulted in one book chapter [Platiša et al., 2012b], one conference proceedings [Platiša et al., 2011a], and several talks and conference abstracts [Platiša et al., 2010a, Platiša et al., 2011b, Platiša et al., 2012a]. A journal article is in preparation [Platiša et al., 2014a].

Several other publications resulted from the contributions to a collaborative work on developing image processing and analysis tools for investigation of the *Ghent Altarpiece*, the polyptych from year 1432 which is considered as one of the most important masterpieces known all over the world: one journal article [Cornelis et al., 2013], one conference proceedings [Ružić et al., 2011], and several publications [Cornelis et al., 2010, Ružić et al., 2010, Cornelis et al., 2011].

The results also attracted attention of wider audience, resulting in several newspaper articles: in the popular Belgian and Dutch science EOS Magazine (June, 2012), in the Flemish newspapers De Standaard[17] (March 27, 2013) and Het Nieuwsblad[18] (March 27, 2013), in the online cultural magazine Cobra of national broadcaster VRT[19] (March 27, 2013), in the Schamper magazine of Ghent University[20] (April 15, 2013). Moreover, the results were mentioned in the Press Release of Ghent University[21] on the occasion of the official start of the physical restoration of the Ghent Altarpiece (September 7, 2012) and the Press Release about the Ghent University research related to the Ghent Altarpiece (March 26, 2013). Finally, the research was pre-

[17] http://www.standaard.be/cnt/dmf20130327_00520016

[18] http://www.nieuwsblad.be/article/detail.aspx?articleid=DMF20130326_00519077

[19] http://www.cobra.be/cm/cobra/kunsten/1.1586571

[20] http://www.schamper.ugent.be/527/op-zoek-naar-het-lam-gods

[21] Persbericht: "Een traditie van innovatief onderzoek van het Lam Gods aan de UGent"

sented in several invited talks at prestigious events: Het Lam Gods Series of Lectures, Provinciaal Administratief Centrum P.A.C. Ghent[22] (November, 2012) and TEDx-Gent[23], Aula, Ghent (June, 2012).

[22] http://www.csct.ugent.be/
[23] http://www.youtube.com/watch?v=kvVb5NG6TLk

7

Concluding remarks

This dissertation studied the problem of evaluating digital images. We developed a range of models and conducted multiple psychovisual studies to assess different kinds of image quality (IQ). In particular, we considered: (1) the usefulness of an image for the task at hand, *i.e.*, the image utility, or the task-based image quality (TaskIQ), (2) the degree of excellence of an image, overall or for specific attributes, *i.e.*, the image beauty, or the technical image quality (TechIQ), and (3) the appearance of an object of interest in an image, *i.e.*, the quality of appearance (ApprIQ). In the following, we briefly review the main contributions of each topical chapter and lay out some of the directions for future research.

We started in **Chapter 2** by presenting the results of two human observer experiments contrasting the main two paradigms of image quality assessment (IQA): the TechIQ and the TaskIQ. At the same time, we assessed the IQ-related effects of common image artifacts (noise, blur) or manipulations (changing display's gamma settings and color saturation, image compression). Interestingly, we found that the agreement between the TechIQ and the TaskIQ may be influenced by the context of the experiment. In particular, the TechIQ ratings which came from a purely technical experimental context disagreed with the TaskIQ ratings which came from the experimental context which involved a clinical task. However, within the clinical context, the TechIQ and the TaskIQ ratings were in agreement.

It would be of great interest to examine these observations in more depth. One obvious next step would involve repeating the same experiments with a larger data set and a larger number of observers. Moreover, it would be of interest to investigate the effects of experimental context for other clinical tasks and especially other imaging modalities. Those findings would be useful for guiding the design of future studies with humans. Another possible track of research would focus on determining thresholds for the maximum level of (attributes of) TechIQ (*i.e.* the minimum level of the corresponding image degradations) which actually bring benefit for the clinical use of certain images. As [Fryback and Thornbury, 1991] point out, "there may be a point beyond which improvement in technical efficacy no longer improves diagnostic accuracy efficacy". That is to say, it would be of importance to define the target levels of

TechIQ for a given application. Those values would serve as benchmarks for the many image processing efforts which insist on improving the performance of the methods in terms of PSNR (see Section 5.8.4) without questioning its real practical benefit.

The work presented in **Chapter 3** and **Chapter 4** focused to the TaskIQ of medical images. Specifically, we considered the task of lesion (signal) detection in the scenario of sequence-browsing image viewing, scrolling through a sequence of image slices. In spite of the growing evidence of the practical diagnostic benefits of volumetric imaging (*e.g.* MRI, CT, DBT), techniques for numerical task-based evaluation of such images are still lacking. Our first contribution in that sense are the two novel designs of model observers, named multi-slice channelized Hotelling observer (msCHO) models. The models are inspired by simplifying assumptions about how the human visual system (HVS) works while browsing through an image sequence: pre-processing the data slice by slice, and then integrating the pre-processed information into a final classification decision (signal-present or signal-absent). Next to our proposed models, we examine another HVS-inspired msCHO design found in the literature. Given the assumptions behind their design, these three models were proposed as candidates for a human-like (anthropomorphic) model observer. In view of that, as our next contribution, we explored and discussed some basic aspects of the practical use of the different model observer designs. The analysis involved model performance for images of different properties (statistical parameters) and model sensitivity to the size of training dataset (in practice, real clinical images are rarely abundant). Apparently, such parameters may significantly affect the predictions of model observer studies. Consequently, it is of utmost importance that they are properly chosen and that the results are interpreted with caution and awareness of the associated limitations.

In practical terms, in Chapter 4 we conducted a series of model observer studies, directly or indirectly aimed at evaluating the utility of medical displays, *i.e.*, to quantify the effects of image display on detectability of the signal in the images. In medical sequence-browsing, one of the major causes of a possible decrease in TaskIQ is the slow response time of a liquid crystal display (LCD). Often, clinicians scroll from one image frame (slice) to the next faster than the LCD luminance change can be physically completed; example LCD response time measurements were provided in Figure 4.9 and Figure 4.14. As a result, the displayed image is often a distorted version of the input one, in which the main difference is in image contrast, a major parameter of detectability. For the purpose of more accurately assessing the effects of the slow LCD displays, we proposed an extension to the msCHO design, the upsampled msCHO model. Unlike the msCHO which considers only the end-of-frame displayed pixel-luminance values, the upsampled msCHO also considers the within-frame pixel-luminance information (see Figure 4.7). Our results demonstrate that integrating the within-frame information into the model observer allows it to be better aware of the LCD temporal luminance variations. Importantly, depending on the details of the luminance changes over time, we found that such models may under- or overestimate signal detectability.

Overall, our results confirm previous findings that the slow temporal response of medical LCDs degrades the detection performance of the observers – the higher the frame rate, the larger the degradation. Undoubtedly, this is a very important recommendation for clinical practice: the rate of browsing through image volumes must be appropriately chosen (not too high) in order to avoid negative effects of the slow LCD temporal response, *i.e.*, in order to avoid introducing degradation in diagnostic accuracy. On the other hand, our msCHO results suggest that the earlier estimates of the extent of these degradations by the conventional CHO model could be overly pessimistic. That is to say, although evidently present, the decrease in signal detectability caused by the slow LCD response time may be not as large and abrupt as previously predicted.

An important confirmation of practical value of the observer models in the process of IQA is the successful use of one of the proposed msCHO models in a preclinical validation of an actual LCD system. Moreover, those msCHO experiments were able to correctly guide the parameters of the followup clinical study with medical specialist observers.

Clearly, a major goal for future research is anthropomorphic models for volumetric images. The human observer study reported at the end of Chapter 4 is already an important step in that direction. There, we examined the effect of image parameters ("task difficulty") on signal detection performance in single-slice (planar) as well as in multi-slice (sequence-browsing) image viewing. Thereby, we aimed to evaluate the factors of the performance differences for 3D versus 2D images. We found that the benefit of 3D is larger for less difficult tasks and smaller for more difficult tasks. These results, together with further research towards understanding the underlying factors of human observer performance (including, but not limited to, contrast sensitivity function (CSF), temporal CSF, masking, internal noise) ought to guide the design modifications to the msCHO model observer such that it can better predict the detection performance of human observers.

Other important directions for future efforts include model improvements towards signal uncertainty, either in terms of unknown signal parameters or unknown signal location, or both. The former aspect is already being addressed in our research group through research led by Prof. Bart Goossens, relying on joint estimation and detection theory. A relevant strategy for developing models of joint signal detection and signal localization could involve eye tracking data of medical specialists and modeling of the HVS search process; somewhat related to the idea of "task-driven attention" in machine learning. Furthermore, it is of interest to develop model observers for new emerging applications such as stereoscopic image viewing. This is also a topic of research in our research group.

In **Chapter 5** we shifted attention to the area of TechIQ assessment. The problem of interest was blur identification in the no-reference scenario (distortion-free image not available). Visually, image blur corresponds to "distorted" edges in the image. Therefore, we aimed at characterizing edges and for that purpose we relied on the

average cone ratio (ACR) of wavelet coefficients; a noise-immune estimate of the local Lipschitz regularity of the signal (see Section 5.3.3). Then, the blur measure was computed as the center of gravity of the histogram of ACR values which correspond to the strongest edges in the image, hence the name CogACR measure. Our experimental results indicated high accuracy of the proposed measure over a wide range of blur levels, even at high levels of noise; this makes it highly competitive to the state-of-the-art.

Furthermore, we examined the effects of image content on the performance of blur measures. As a descriptor of image content, we proposed using the histogram of ACR values (HistACR) corresponding to the dominant edges in the image. Moreover, we proposed a novel HistACR-based measure of image similarity. While existing similarity measures are often context-based, our technique quantifies the similarity of edges in the images. The measure was able to successfully identify images with similar behavior in varying blur, not only according to the mathematical parameters but also according to the humans. This suggests potential for advancing the proposed set of measures towards content-aware assessment of image blurriness. Thus, future efforts may be directed at refining the methodology for content classification for the purpose of IQA. One potential direction may concern using existing classification techniques such as the popular support vector machine.

In addition, we observed that the CogACR is sensitive to image content in the way which intuitively corresponds to the sensitivity of the HVS, higher sensitivity to small distortions in high frequency image content compared to the low frequency one. Accordingly, we could further investigate dominant parameters of human perception of blurriness (through psychovisual studies) and use those findings to adjust the algorithms to better predict humans. We have already started preliminary investigations in that direction.

Finally, the same as in the case of model observers, it seems worthwhile to explore the emerging trend of incorporating image saliency information in the techniques for IQA. Moving in that direction, it would be of interest to explore several alternatives; for example, applying the measure only to the model-predicted salient regions in the image (rather than on the whole image area as we do now), or using the saliency information as a weighting factor for individual ACR coefficients, or even combining the saliency information with the aforementioned dominant parameters of human perception of blurriness.

The thesis ends with the work of **Chapter 6** which focused on developing methods for quantifying attributes of ApprIQ of pearls and pearl-like objects in digital images of art paintings. Because the surface reflectance is among the most notable characteristics of jewels (also in paintings), it was essential to have spatial information involved in the analysis of pearl images. To do that, we choose to work with the so-called image spatiogram, the image histogram with added spatial information about the histogram bins. First, we proposed a method for visualizing the multidimensional spatiogram data; the problem which has not been addressed before. Next, we

evaluated the performance of an existing spatiogram similarity measure suggested by the literature. While the existing measure agreed fairly well with the humans in terms of how it ranked the pearls based on similarity, the major drawback was the lack of details about the measured dissimilarities. As an illustration, the measure could not indicate if the pearls differed in the smoothness of their surfaces, or in the size of their highlight areas. These details, however, are a very important aspect for art history analysis. To address that problem, we developed a (image restoration) method for matching spatiograms of different images. This was used as a toll for our explorative analysis of the relationship between the dominant factors of the appearance of pearls (and pearl-like objects) in images and the properties of the corresponding spatiograms. Based on those investigations, we proposed a set of novel spatiogram-based measures which quantify selected features of pearl appearance; mainly, the appearance of surface smoothness and several aspects regarding object symmetry. The methods were evaluated on a range of pearls and beads, both painted and photographed. Overall, the observed agreement between the new measures and the visually observed image features makes the proposed approach a promising candidate for practical use in characterizing pearls in paintings.

Clearly, for a more comprehensive characterization of the object appearance, the set of measures shall be extended to allow a more precise and more detailed description of the relevant features. For example, as suggested by the questionnaire results from our human observer study, one possibly important direction for the future investigations could be characterization of the appearance of the highlight area of the pearl (referring to its position, shape, edges, and maybe other properties of the highlight). Likewise, future human experiments should look deeper into the specific attributes of ApprIQ; previously, our experiments involved only the rating of overall similarity of pearls. Tentative applications for the proposed techniques and their advances include assisting art historians in better understanding the differences or similarities between different artists and their ways of painting pearls, as well as artist identification. Beyond the domain of artwork analysis, these kinds of techniques could be applied in the area of dermatology imaging, for example, to characterize the appearance of a skin lesion. Likely, the exact attributes of appearance may need to be revised and possibly redefined but the core idea of the approach remains the same.

Finally, our methods so far do not explicitly take into account the color information in images. Given that the color is now becoming present not only in commercial imaging applications (*e.g.* digital cameras and television), but also in scientific imaging (*e.g.* digital pathology imaging discussed in Chapter 2), an important direction for future work concerns extending the proposed models to color images, including all TaskIQ, TechIQ, and ApprIQ related methods developed in this dissertation.

A

List of publications

A.1 Publications in international journals

1. Platiša, L., Goossens, B., Vansteenkiste, E., Park, S., Gallas, B. D., Badano, A., and Philips, W. (2011e). Channelized Hotelling observers for the assessment of volumetric imaging data sets. *J. Opt. Soc. Am. A*, 28(6):1145–1163

2. Cornelis, B., Ruzic, T., Gezels, E., Dooms, A., Pižurica, A., Platiša, L., Cornelis, J., Martens, M., De Mey, M., and Daubechies, I. (2013). Crack detection and inpainting for virtual restoration of paintings: The case of the Ghent Altarpiece. *Signal Process.*, 93(3, SI):605–619

3. Platiša, L., Kumcu, A., Platiša, M., Vansteenkiste, E., Gallas, B. D., Deblaere, K., Badano, A., and Philips, W. (2014b). Lesion detection performance in single- versus multi-slice image readings: results from human and model observer studies. (In preparation)

4. Platiša, L. and Pižurica, A. (2014). No-reference blur estimation based on the average cone ratio in the wavelet domain. (In preparation)

5. Platiša, L., Cornelis, B., Ružić, T., Pižurica, A., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2014a). Spatiogram-based descriptors for quality of appearance of pearl-like objects in the images. (In preparation)

A.2 Book chapters

1. Platiša, L., Cornelis, B., Ružić, T., Pižurica, A., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2012b). *Vision and material : interaction between art and science in Jan Van Eyck's time*, chapter Spatiogram Features to Characterize Pearls and Beads and other Small Ball-shaped Objects in Paintings, pages 315–329. KVAB PRESS

2. Pižurica, A., Platiša, L., Ružić, T., Cornelis, B., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2013). *Het Lam Gods Series of Lectures*, chapter Virtual Restoration and Mathematical Analysis of Pearls in the Adoration of the Mystic Lamb. (To appear)

A.3 Publications in international and national conferences

1. Ortiz Jaramillo, B., Kumcu, A., Platia, L., and Philips, W. (2014). A full reference video quality measure based on motion differences and saliency maps evaluation. In *Proc. 9th International Conference on Computer Vision Theory and Applications*, Lisbon, Portugal. (In press)

2. Kumcu, A. E., Bombeke, K., Chen, H., Jovanov, L., Platiša, L., Luong, Q., Van Looy, J., Van Nieuwenhove, Y., Schelkens, P., and Philips, W. (2014). Visual quality assessment of H.264/AVC compressed laparoscopy video. In *Proc. SPIE Medical Imaging*. (In press)

3. Rousson, J., Couturou, J., Vetsuypens, A., Platiša, L., Kumcu, A., Kimpe, T., and Philips, W. (2014). Subjective quality and depth assessment in stereoscopic viewing of volume-rendered medical images. In *Proc. SPIE Electronic Imaging*. (In press)

4. Platiša, L., Van Brantegem, L., Vander Haeghen, Y., Marchessoux, C., Vansteenkiste, E., and Philips, W. (2013b). Psycho-visual evaluation of image quality attributes in digital pathology slides viewed on a medical color LCD display. In *Proc. SPIE Medical Imaging*, volume 8676, page 86760J, Orlando, Florida, USA

5. Goossens, B., Luong, Q., Platiša, L., and Philips, W. (2013). Objectively measuring signal detectability, contrast, blur and noise in medical images using channelized joint observers. In *Proc. SPIE Medical Imaging*, page 11

6. Platiša, L., Kumcu, A., Platiša, M., Vansteenkiste, E., Deblaere, K., Badano, A., and Philips, W. (2012c). Volumetric detection tasks with varying complexity: human observer performance. In Abbey, C. K. and Mello-Thoms, C. R., editors, *Proc. SPIE Medical Imaging*, volume 8318, page 83180S

7. Goossens, B., Platiša, L., and Philips, W. (2012b). Theoretical performance analysis of multislice channelized Hotelling observers. In Abbey, C. K. and Mello-Thoms, C. R., editors, *Proc. SPIE Medical Imaging*, volume 8318, page 83180U

8. Kumcu, A., Platiša, L., Platiša, M., Vansteenkiste, E., Deblaere, K., Badano, A., and Philips, W. (2012b). Reader behavior in a detection task using single- and

multislice image datasets. In Abbey, C. K. and Mello-Thoms, C. R., editors, *Proc. SPIE Medical Imaging*, volume 8318, page 831803

9. Luong, Q., Goossens, B., Aelterman, J., Platiša, L., and Philips, W. (2012). Optimizing image quality in MRI: on the evaluation of k-space trajectories for under-sampled MR acquisition. In *Proc. IEEE International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 25–26

10. Goossens, B., Luong, Q., Platiša, L., and Philips, W. (2012a). Optimizing image quality using test signals: trading off blur, noise and contrast. In *Proc. IEEE International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 260–265

11. Platiša, L., Cornelis, B., Ružić, T., Pižurica, A., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2011a). Spatiogram features to characterize pearls in paintings. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 801–804

12. Platiša, L., Pizurica, A., Vansteenkiste, E., and Philips, W. (2011j). No-reference blur estimation based on the average cone ratio in the wavelet domain. In Akopian, D., Creutzburg, R., Snoek, C. G. M., Sebe, N., and Kennedy, L., editors, *Proc. SPIE Electronic Imaging*, volume 7881, page 78811D

13. Platiša, L., Marchessoux, C., Goossens, B., and Philips, W. (2011g). Performance evaluation of medical LCD displays using 3D channelized Hotelling observers. In Manning, D. J. and Abbey, C. K., editors, *Proc. SPIE Medical Imaging*, volume 7966, page 79660T

14. Platiša, L., Marchessoux, C., Kimpe, T., Vansteenkiste, E., Badano, A., and Philips, W. (2011h). Channelized Hotelling observers for signal detection in stack-mode reading of volumetric images on medical displays with slow response time. In *Proc. IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pages 2697–2702

15. Despotović, I., Segers, I., Platiša, L., Vansteenkiste, E., Pižurica, A., Deblaere, K., and Philips, W. (2011a). Automatic 3D graph cuts for brain cortex segmentation in patients with focal cortical dysplasia. In *Proc. IEEE Conference on Engineering in Medicine and Biology society (EMBC)*, pages 7981–7984

16. Ružić, T., Cornelis, B., Platiša, L., Pižurica, A., Dooms, A., Philips, W., Martens, M., De Mey, M., and Daubechies, I. (2011). Virtual restoration of the Ghent altarpiece using crack detection and inpainting. In Blanc-Talon, J., Kleihorst, R., Philips, W., Popescu, D., and Scheunders, P., editors, *Lecture Notes in Computer Science*, volume 6915, pages 417–428

17. Platiša, L., De Smet, A., Despotović, I., Kumcu, A., Vansteenkiste, E., Deblaere, K., Pizurica, A., and Philips, W. (2011d). Measuring cortical thickness in brain MRI volumes to detect focal cortical dysplasia (FCD) in epilepsy patients. In *Proc. International Society for Magnetic Resonance in Medicine (ISMRM)*, number 19, pages 2438–2438

18. Despotović, I., Segers, I., Platiša, L., Vansteenkiste, E., Pižurica, A., Deblaere, K., and Philips, W. (2011b). Brain MRI segmentation for focal cortical dysplasia lesion detection. In *Proc. International Society for Magnetic Resonance in Medicine (ISMRM)*, number 19, pages 4277–4277

19. Platiša, L., Goossens, B., Vansteenkiste, E., Badano, A., and Philips, W. (2010c). Using channelized Hotelling observers to quantify temporal effects of medical liquid crystal displays on detection performance. In Manning, D. J. and Abbey, C. K., editors, *Proc. SPIE Medical Imaging*, volume 7627, page 76270U

20. Goossens, B., Platiša, L., Vansteenkiste, E., and Philips, W. (2010). The use of steerable channels for detecting asymmetrical signals with random orientations. In Manning, D. J. and Abbey, C. K., editors, *Proc. SPIE Medical Imaging*, volume 7627, page 76270S

21. Lukić, N., Platiša, L., Pižurica, A., Philips, W., and Temerinac, M. (2010). Real-time wavelet based blur estimation on cell BE platform. In Truchetet, F. and Laligant, O., editors, *Proc. SPIE Electronic Imaging*, volume 7535, page 75350C

22. Platiša, L., Goossens, B., Vansteenkiste, E., Badano, A., and Philips, W. (2009b). Channelized Hotelling observers for the detection of 2D signals in 3D simulated images. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 1761–1764

23. Platiša, L., Pižurica, A., Vansteenkiste, E., and Philips, W. (2009c). Image blur estimation based on the average cone of ratio in the wavelet domain. In Truchetet, F. and Laligant, O., editors, *Proc. SPIE Electronic Imaging*, volume 7248, page 10

24. Platiša, L., Vansteenkiste, E., Goossens, B., Marchessoux, C., Kimpe, T., and Philips, W. (2009d). Optimization of medical imaging display systems: using the channelized Hotelling observer for detecting lung nodules: experimental study. In Sahiner, B. and Manning, D. J., editors, *Proc. SPIE Medical Imaging*, volume 7263, page 72630P

A.4 Abstracts in international and national conferences

1. Platiša, L., Van Brantegem, L., Kumcu, A., Marchessoux, C., Vansteenkiste, E., and Philips, W. (2013a). Effects of common image manipulations on diagnostic

performance in digital pathology human study. In *Conference of Medical Image Perception Society (MIPS), Abstracts*, Washington, DC, USA

2. Maidment, A. and Platiša, L. (2013). The roles and limitations of model observer studies. In *Conference of Medical Image Perception Society (MIPS), Abstracts*, Washington, DC, USA

3. Kumcu, A., Elprama, S., Vermeulen, L., Duysburgh, P., Platiša, L., VanNieuwenhove, Y., Van De Winkel, N., Jacobs, A., Van Looy, J., and Philips, W. (2013). Effect of video latency on performance and subjective experience in laparoscopic surgery. In *Conference of Medical Image Perception Society (MIPS), Abstracts*, Washington, DC, USA

4. Qu, X., Kumcu, A., Platiša, L., Despotovic, I., Deblaere, K., Bai, T., and Philips, W. (2013). Blur estimation at the gray-white matter boundary for focal cortical dysplasia in magnetic resonance imaging. In *IEEE Annual International Conference of the Engineering in Medicine and Biology Society (EMBS), Abstracts*

5. Qu, X., Platiša, L., Despotovic, I., Kumcu, A., Deblaere, K., Bai, T., and Philips, W. (2014). Automatic brain atlas in magnetic resonance image for focal cortical dysplasia patients. In *IEEE International Symposium on Biomedical Imaging (ISBI), Abstracts*. (In press)

6. Platiša, L., Cornelis, B., Ružić, T., Pižurica, A., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2012a). Spatiogram features to characterize pearls in the Ghent Altarpiece. In *Symposium for the Study of Underdrawing and Technology in Painting, Abstracts*. Ghent University, Department of Telecommunications and information processing

7. Kumcu, A., Platiša, L., Despotovic, I., Vansteenkiste, E., Pizurica, A., Deblaere, K., and Philips, W. (2012a). Multi-modal measurement of cortical thickness in brain MRI for Focal Cortical Dysplasia detection. In *Annual meeting of the International Society for Magnetic Resonance in Medicine (ISMRM) Benelux Chapter, Abstracts*

8. Platiša, L., Pižurica, A., Vansteenkiste, E., and Philips, W. (2011i). No-reference blur estimation based on the average cone ratio in the wavelet domain. In *IEEE International Workshop on Quality of Multimedia Experience (QoMEX), Abstracts*

9. Platiša, L., Kumcu, A., Platiša, M., Vansteenkiste, E., Gallas, B., Deblaere, K., Badano, A., and Philips, W. (2011f). Model and human observers studies in volumetric images for detection tasks with varying complexity. In *Conference of Medical Image Perception Society (MIPS), Abstracts*, pages 27–27

10. Platiša, L., Cornelis, B., Ružić, T., Pižurica, A., Dooms, A., Schelkens, P., Martens, M., De Mey, M., and Daubechies, I. (2011b). Spatiogram features to characterize pearls in paintings. In *International workshop on Image Processing for Art Investigation, Abstracts*

11. Platiša, L., De Smet, A., Despotović, I., Kumcu, A., Vansteenkiste, E., Deblaere, K., Pižurica, A., and Philips, W. (2011c). Measuring cortical thickness in brain MRI volumes to detect focal cortical dysplasia (FCD) in epilepsy patients. In *Annual meeting of the International Society for Magnetic Resonance in Medicine (ISMRM) Benelux Chapter, Abstracts*

12. Kumcu, A., Platiša, L., Goossens, B., and Philips, W. (2011a). Subjective and objective quality evaluation of compressed medical video sequences. In *IEEE International Workshop on Quality of Multimedia Experience (QoMEX), Abstracts*

13. Kumcu, A., Platiša, L., Platiša, M., Vansteenkiste, E., Deblaere, K., Badano, A., and Philips, W. (2011c). Trends in reader behavior for a signal detection task in multi- and single-slice images. In *Conference of Medical Image Perception Society (MIPS), Abstracts*

14. Kumcu, A., Platiša, L., Goossens, B., and Philips, W. (2011b). Visual quality assessment of H.264 and motion JPEG compressed laparoscopy video. In *IEEE Engineering in Medicine and Biology Society (EMBS) Benelux Chapter, Annual symposium, Abstracts*

15. Cornelis, B., Ruzic, T., Gezels, E., Dooms, A., Pižurica, A., Platiša, L., Martens, M., Schelkens, P., De Mey, M., and Daubechies, I. (2011). Crack detection and inpainting for virtual restoration of paintings: the case of the Ghent altarpiece. In *International workshop on Image Processing for Art Investigation, Abstracts*

16. Despotović, I., Segers, I., Platiša, L., Vansteenkiste, E., Pižurica, A., Deblaere, K., and Philips, W. (2011c). Brain MRI segmentation for focal cortical dysplasia lesion detection. In *Annual meeting of the International Society for Magnetic Resonance in Medicine (ISMRM) Benelux Chapter, Abstracts*

17. Vetsuypens, A., Besnehard, Q., Platiša, L., Arnault, E., Marchessoux, C., Kimpe, T., Xthona, A., and Philips, W. (2011). Creating a modality-optimized medical display for DBT based on MEVIC simulations. In *Conference of Medical Image Perception Society (MIPS), Abstracts*

18. Platiša, L., Lukić, N., Pižurica, A., Vansteenkiste, E., and Philips, W. (2010d). Image blur estimation based on the average cone of ratio in the wavelet domain. In *Sparsity and modern mathematical methods for high dimensional data, Abstracts*, pages 24–24

19. Platiša, L. (2010). No-reference wavelet-based blur metric for image quality assessment. In *UGent-FirW Doctoraatssymposium, Abstracts*, pages 152–152. Universiteit Gent. Faculteit Ingenieurswetenschappen

20. Platiša, L., Cornelis, B., Ružić, T., Pižurica, A., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2010a). Pearls and beads in Jan van Eyck's paintings. In *Vision and material : interaction between art and science in Jan Van Eyck's time, Abstracts*

21. Platiša, L., Goossens, B., Vansteenkiste, E., Badano, A., and Philips, W. (2010b). Quantifying temporal effects of medical LCD monitors on lesion detectability. In *Belgian Day on Biomedical Engineering, 9th, Abstracts*

22. Cornelis, B., Platiša, L., Ruzic, T., Dooms, A., Pižurica, A., Martens, M., De Mey, M., and Daubechies, I. (2010). Teaching a computer about shapes in paintings. In *Vision and material : interaction between art and science in Jan Van Eyck's time, Abstracts*

23. Ružić, T., Cornelis, B., Platiša, L., Pižurica, A., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2010). Craquelure inpainting in art work. In *Vision and material : interaction between art and science in Jan Van Eyck's time, Abstracts*

24. Platiša, L., Goossens, B., Vansteenkiste, E., Badano, A., and Philips, W. (2009a). Channelized Hotelling observers for detection tasks in multi-slice images. In *Conference of Medical Image Perception Society (MIPS), Abstracts*, pages 36–36

25. Platiša, L. (2008). Optimization of medical imaging display systems using the channelized Hotelling observer. In *UGent-FirW Doctoraatssymposium, Abstracts*, pages 194–195

Bibliography

[Abbey and Barrett, 2001] Abbey, C. K. and Barrett, H. H. (2001). Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability. *J. Opt. Soc. Am. A*, 18(3):473–488.

[A.C.R.Committee, 1999] A.C.R.Committee (1999). Mammography quality control manual.

[Ahumada and Null, 1993] Ahumada, A. and Null, C. H. (1993). Image quality: A multidimensional problem. *Digital images and human vision*, pages 141–148.

[American Association of Physicists in Medicine, 1993] American Association of Physicists in Medicine (1993). Specification and acceptance testing of computed tomography scanners: report no. 39. Technical Report 39, New York, NY.

[Andersson et al., 2008] Andersson, I., Ikeda, D. M., Zackrisson, S., Ruschin, M., Svahn, T., Timberg, P., and Tingberg, A. (2008). Breast tomosynthesis and digital mammography: a comparison of breast cancer visibility and BIRADS classification in a population of cancers with subtle mammographic findings. *Eur. Radiol.* , 18:2817–2825.

[Autin et al., 1999] Autin, M. C., Gonzalez-Palacios, A., and Scarisbrick, D. (1999). *Jewels in Painting*. Skira, Milano, Italy.

[Avanaki et al., 2013] Avanaki, A. N., Espig, K. S., Marchessoux, C., Krupinski, E. A., Bakic, P. R., Kimpe, T. R., and Maidment, A. D. (2013). Integration of spatio-temporal contrast sensitivity with a multi-slice channelized Hotelling observer. In *Proc. SPIE Medical Imaging*, pages 86730H–86730H. International Society for Optics and Photonics.

[Badano, 2009] Badano, A. (2009). Effect of slow display on detectability when browsing large image datasets. *J. Soc. Inf. Disp.*, 17(11):891–896.

[Badano et al., 2003] Badano, A., Flynn, M. J., Martin, S., and Kanicki, J. (2003). Angular dependence of the luminance and contrast in medical monochrome liquid crystal displays. *Med. Phys.*, 30:2602.

[Badano et al., 2004] Badano, A., Gagne, R. M., Jennings, R. J., Drilling, S. E., Imhoff, B. R., and Muka, E. (2004). Noise in flat-panel displays with subpixel structure. *Med. Phys.*, 31:715.

[Badano and Gallas, 2006] Badano, A. and Gallas, B. D. (2006). Detectability decreases with off-normal viewing in medical liquid crystal displays. *Acad. Radiol.*, 13(2):210–218.

[Bao et al., 2005] Bao, P., Zhang, L., and Wu, X. (2005). Canny edge detection enhancement by scale multiplication. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(9):1485–1490.

[Barrett, 1990] Barrett, H. H. (1990). Objective assessment of image quality: effects of quantum noise and object variability. *J. Opt. Soc. Am. A*, 7(7):1266–1278.

[Barrett et al., 1998] Barrett, H. H., Abbey, C. K., and Clarkson, E. (1998). Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions. *J. Opt. Soc. Am. A*, 15(6):1520–1535.

[Barrett et al., 1995] Barrett, H. H., Denny, J., Wagner, R. F., and Myers, K. J. (1995). Objective assessment of image quality. ii. fisher information, fourier crosstalk, and figures of merit for task performance. *J. Opt. Soc. Am. A*, 12(5):834–852.

[Barrett and Myers, 2004] Barrett, H. H. and Myers, K. J. (2004). *Foundations of Image Science*. John Wiley and Sons, New York.

[Barrett et al., 2006] Barrett, H. H., Myers, K. J., Devaney, N., and Dainty, C. (2006). Objective assessment of image quality. iv. application to adaptive optics. *J. Opt. Soc. Am. A*, 23(12):3080–3105.

[Barrett et al., 2001] Barrett, H. H., Myers, K. J., Gallas, B. D., Clarkson, E., and Zhang, H. (2001). Megalopinakophobia: its symptoms and cures. In Antonuk, L. E. and Yaffe, M. J., editors, *Proc. SPIE Medical Imaging*, volume 4320, pages 299–307. SPIE.

[Barrett et al., 1993] Barrett, H. H., Yao, J., Rolland, J. P., and Myers, K. J. (1993). Model observers for assessment of image quality. *Proceedings of the National Academy of Sciences of the United States of America*, 90(21):9758.

[Barten, 1999] Barten, P. G. J. (1999). *Contrast Sensitivity of the Human Eye and its Effects on Image Quality*. SPIE Press, Bellingham, WA.

[Beiden et al., 2000] Beiden, S. V., Wagner, R. F., and Campbell, G. (2000). Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. *Acad. Radiol.*, 7(5):341–349.

[Birchfield and Rangarajan, 2005] Birchfield, S. T. and Rangarajan, S. (2005). Spatiograms versus histograms for region-based tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1158–1163.

[Bochud et al., 1999] Bochud, F. O., Abbey, C. K., and Eckstein, M. P. (1999). Statistical texture synthesis of mammographic images with clustered lumpy backgrounds. *Opt. Express*, 4:33–43.

[Bongartz et al., 2004] Bongartz, G., Golding, S. J., Jurik, A. G., Leonardi, M., Van Persijn van Meerten, E., Rodríguez, R., Schneider, K., Calzado, A., Geleijns, J., Jessen, K. A., Panzer, W., Shrimpton, P. C., and Tosi, G. (2004). European guidelines for multislice computed tomography. *European Commission*.

[Brankov, 2011] Brankov, J. (2011). Optimization of the internal noise models for channelized Hotelling observer. In *Proc. IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1788 –1791.

[Bunch et al., 1977] Bunch, P. C., Hamilton, J. F., Sanderson, G. K., and Simmons, A. H. (1977). A free response approach to the measurement and characterization of radiographic observer performance. In *Application of Optical Instrumentation in Medicine*, pages 124–135. International Society for Optics and Photonics.

[Burgess, 1995a] Burgess, A. (1995a). Image quality, the ideal observer, and human performance of radiologic decision tasks. *Acad. Radiol.*, 2(6):522 – 526.

[Burgess, 1995b] Burgess, A. E. (1995b). Comparison of receiver operating characteristic and forced choice observer performance measurement methods. *Med. Phys.*, 22(5):643–655.

[Burgess, 1999a] Burgess, A. E. (1999a). Visual signal detection with two-component noise: low-pass spectrum effects. *J. Opt. Soc. Am. A*, 16(3):694–704.

[Burgess, 1999b] Burgess, A. E. (1999b). Visual signal detection with two-component noise: low-pass spectrum effects. *J. Opt. Soc. Am. A*, 16(3):694–704.

[Burgess et al., 2001] Burgess, A. E., Jacobson, F. L., and Judy, P. F. (2001). Human observer detection experiments with mammograms and power-law noise. *Med. Phys.*, 28(4):419–437.

[Burgess et al., 1982] Burgess, A. E., Wagner, R. F., and Jennings, R. J. (1982). Human signal detection performance for noisy medical images. In *Proc. SPIE Physics and Engineering in Medical Imaging*, volume 0372, pages 99–105.

[Cai et al., 2012] Cai, J.-F., Ji, H., Liu, C., and Shen, Z. (2012). Framelet-based blind motion deblurring from a single image. *IEEE Trans. Image Process.*, 21(2):562 – 572.

[Candès et al., 2006] Candès, E., Demanet, L., Donoho, D., and Ying, L. (2006). Fast Discrete Curvelet Transforms. *Multiscale modeling and simulation*, 5(3):861–899.

[Castella et al., 2009] Castella, C., Eckstein, M. P., Abbey, C. K., Kinkel, K., Verdun, F. R., Saunders, R. S., Samei, E., and Bochud, F. O. (2009). Mass detection on mammograms: influence of signal shape uncertainty on human and model observers. *J. Opt. Soc. Am. A*, 26(2):425–436.

[Castella et al., 2008] Castella, C., Kinkel, K., Descombes, F., Eckstein, M. P., Sottas, P.-E., Verdun, F. R., and Bochud, F. O. (2008). Mammographic texture synthesis: second-generationclustered lumpy backgrounds using agenetic algorithm. *Opt. Express*, 16(11):7595–7607.

[Cavaro-Ménard et al., 2013] Cavaro-Ménard, C., Le Callet, P., Hunault, G., Pépion, R., and Rousselet, M.-C. (2013). Effect of JPEG2000 Compression on the Visual Quality of Virtual Liver Biopsy Slides: Subjective and Objective Measurements. In *Conference of Medical Image Perception Society (MIPS), Abstracts*, Washington, DC, USA. Medical Image Perception Society (MIPS).

[Cha, 2000] Cha, S.-H. (2000). Efficient algorithms for image template and dictionary matching. *J. Math. Imaging Vis.*, 12(1):81–90.

[Chakraborty, 2010] Chakraborty, D. (2010). *The handbook of medical image perception and techniques*, chapter The role of expertise in radiologic image interpretation, pages 216–239. Cambridge University Press, New York.

[Chakraborty, 1989] Chakraborty, D. P. (1989). Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. *Med. Phys.*, 16(4):561–568.

[Chakraborty, 2006] Chakraborty, D. P. (2006). Analysis of location specific observer performance data: Validated extensions of the Jackknife Free-Response (JAFROC) method. *Acad. Radiol.*, 13(10):1187 – 1193.

[Chakraborty and Berbaum, 2004] Chakraborty, D. P. and Berbaum, K. S. (2004). Observer studies involving detection and localization: modeling, analysis, and validation. *Med. Phys.*, 31:2313.

[Chawla et al., 2007] Chawla, A. S., Samei, E., Saunders, R., Abbey, C., and Delong, D. (2007). Effect of dose reduction on the detection of mammographic lesions: A mathematical observer model analysis. *Med. Phys.*, 34:3385.

[Chawla et al., 2008] Chawla, A. S., Samei, E., Saunders, R. S., Lo, J. Y., and Baker, J. A. (2008). A mathematical model platform for optimizing a multiprojection breast imaging system. *Med. Phys.*, 35:1337.

[Chen and Barrett, 2005] Chen, L. and Barrett, H. H. (2005). Task-based lens design with application to digital mammography. *J. Opt. Soc. Am. A*, 22(1):148–167.

[Chen et al., 2002] Chen, M., Bowsher, J., Baydush, A., Gilland, K., DeLong, D., and Jaszczak, R. (2002). Using the Hotelling observer on multislice and multiview simulated SPECT myocardial images. *IEEE Trans. Nucl. Sci.*, 49(3):661 – 667.

[Chen, 2012] Chen, Z. (2012). Object-based attention: A tutorial review. *Attention Perception & Psychophysics*, 74(5):784–802.

[Chikkerur et al., 2011] Chikkerur, S., Sundaram, V., Reisslein, M., and Karam, L. (2011). Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Trans. Broadcast.*, 57(2):165–182.

[Clarkson et al., 2006] Clarkson, E., Kupinski, M. A., and Barrett, H. H. (2006). A probabilistic development of the MRMC method. *Acad. Radiol.*, 13 (10):1410–1421.

[Coifman and Donoho, 1995] Coifman, R. and Donoho, D. L. (1995). *Wavelets and Statistics*, chapter Translation Invariant Denoising. Springer, New York, NY, USA.

[Conaire et al., 2007] Conaire, C. O., O'Connor, N. E., and Smeaton, A. F. (2007). An improved spatiogram similarity measure for robust object localisation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages I–1069 –I–1072.

[Cornelis et al., 2010] Cornelis, B., Platiša, L., Ruzic, T., Dooms, A., Pižurica, A., Martens, M., De Mey, M., and Daubechies, I. (2010). Teaching a computer about shapes in paintings. In *Vision and material : interaction between art and science in Jan Van Eyck's time, Abstracts*.

[Cornelis et al., 2013] Cornelis, B., Ruzic, T., Gezels, E., Dooms, A., Pižurica, A., Platiša, L., Cornelis, J., Martens, M., De Mey, M., and Daubechies, I. (2013). Crack detection and inpainting for virtual restoration of paintings: The case of the Ghent Altarpiece. *Signal Process.*, 93(3, SI):605–619.

[Cornelis et al., 2011] Cornelis, B., Ruzic, T., Gezels, E., Dooms, A., Pižurica, A., Platiša, L., Martens, M., Schelkens, P., De Mey, M., and Daubechies, I. (2011). Crack detection and inpainting for virtual restoration of paintings: the case of the Ghent altarpiece. In *International workshop on Image Processing for Art Investigation, Abstracts*.

[Cornsweet, 1962] Cornsweet, T. N. (1962). The staircase-method in psychophysics. *Am. J. Psychol.*, 75:485–491.

[Daubechies, 1988] Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 41(7):909–996.

[Daubechies, 1992] Daubechies, I. (1992). *Ten Lectures on Wavelets*. Number 61 in CBMS/NSF Series in Applied Math.

[De Mey, 2008] De Mey, M. (2008). *Jan van Eyck and the Representation of Glow, in: Anna De Floriani & Maria Clelia Galassi (eds.) Culture figurative a confronto tra Fiandre e Italia dal XV al XVII secolo*. Silvana Editoriale, Milano.

[Despotović et al., 2011a] Despotović, I., Segers, I., Platiša, L., Vansteenkiste, E., Pižurica, A., Deblaere, K., and Philips, W. (2011a). Automatic 3D graph cuts for brain cortex segmentation in patients with focal cortical dysplasia. In *Proc. IEEE Conference on Engineering in Medicine and Biology society (EMBC)*, pages 7981–7984.

[Despotović et al., 2011b] Despotović, I., Segers, I., Platiša, L., Vansteenkiste, E., Pižurica, A., Deblaere, K., and Philips, W. (2011b). Brain MRI segmentation for focal cortical dysplasia lesion detection. In *Proc. International Society for Magnetic Resonance in Medicine (ISMRM)*, number 19, pages 4277–4277.

[Despotović et al., 2011c] Despotović, I., Segers, I., Platiša, L., Vansteenkiste, E., Pižurica, A., Deblaere, K., and Philips, W. (2011c). Brain MRI segmentation for focal cortical dysplasia lesion detection. In *Annual meeting of the International Society for Magnetic Resonance in Medicine (ISMRM) Benelux Chapter, Abstracts*.

[Diaz et al., 2011] Diaz, I., Timberg, P., Zhang, S., Abbey, C., Verdun, F., and Bochud, F. O. (2011). Development of model observers applied to 3D breast tomosynthesis microcalcifications and masses. In *Proc. SPIE Medical Imaging*, volume 7966, page 79660F. SPIE.

[DICOM, 2009] DICOM (2009). DICOM Standard Committee, Working Groups 26, Pathology, Digital Imaging and Communications in Medicine (DICOM) Supplement 145: Whole Slide Microscopic Image IOD and SOP Classes.

[Donoho, 1999] Donoho, D. (1999). Wedgelets: Nearly minimax estimation of edges. *Annals of Statistics*, 27(3):859–897.

[Dorfman et al., 1992] Dorfman, D., Berbaum, K., and Metz, C. (1992). Receiver Operating Characteristic Rating Analysis: Generalization to the Population of Readers and Patients with the Jackknife Method. *Invest. Radiol.*, 27:723–731.

[Ducottet et al., 2004] Ducottet, C., Fournel, T., and Barat, C. (2004). Scale-adaptive detection and local characterization of edges based on wavelet transform. *Signal Process.*, 84(11):2115–2137.

[Duda and Hart, 1972] Duda, R. O. and Hart, P. E. (1972). Use of the Hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15:11–15.

[Eckstein et al., 1998] Eckstein, M. P., Abbey, C. K., and Whiting, J. S. (1998). Human vs model observers in anatomic backgrounds. In *Proc. SPIE Medical Imaging*, pages 16–26. SPIE.

[Eckstein et al., 1997] Eckstein, M. P., Ahumada, A. J., and Watson, A. B. (1997). Visual signal detection in structured backgrounds. II. Effects of contrast gain control, background variations, and white noise. *J. Opt. Soc. Am. A*, 14(9):2406–2419.

[Elder and Zucker, 1998] Elder, J. and Zucker, S. (1998). Local scale control for edge detection and blur estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(7):699–716.

[Engeldrum, 2004] Engeldrum, P. (2004). A theory of image quality: The Image Quality Circle. *J. Imaging Sci. Technol.*, 48(5):447–457.

[Engelke et al., 2011] Engelke, U., Kaprykowsky, H., Zepernick, H.-J., and Ndjiki-Nya, P. (2011). Visual attention in quality assessment. *IEEE Signal Process. Mag.*, 28(6):50 –59.

[EUCommission, 1996a] EUCommission (1996a). European guidelines on quality criteria for diagnostic radiographic images. European Commission.

[EUCommission, 1996b] EUCommission (1996b). European guidelines on quality criteria for diagnostic radiographic images in pediatrics. European Commission.

[Farn, 1986] Farn, A. E. (1986). *Pearls, Natural, Cultured and Imitation*. London, Butterworths.

[Farnand et al., 2013] Farnand, S., Jiang, J., and Frey, F. (2013). Current practices in fine art reproduction: Project summary. In *Archiving 2013*, pages 48–53, Washington, DC.

[Farnand et al., 2009] Farnand, S. P., Frey, F. S., and Anderson, E. (2009). Benchmarking art image interchange cycles: Image quality experimentation. In *Congress of the International Color Association (AIC)*, Sydney, New South Wales, Australia. International Colour Association.

[Farnsworth, 1947] Farnsworth, D. (1947). *The Farnsworth Dichotomous Test for Color Blindness: Panel D-15*. Psychological Corporation.

[Ferzli and Karam, 2005] Ferzli, R. and Karam, L. (2005). No-reference objective wavelet based noise immune image sharpness metric. In *Proc. IEEE International Conference on Image Processing (ICIP)*, volume 1-5, pages 1157–1160. Genoa, ITALY, SEP 11-14, 2005.

[Ferzli and Karam, 2009] Ferzli, R. and Karam, L. (2009). A No-Reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur (JNB). *IEEE Trans. Image Process.*, 18(4):717 –728.

[Ferzli et al., 2005] Ferzli, R., Karam, L. J., and Caviedes, J. (2005). A robust image sharpness metric based on kurtosis measurement of wavelet coefficients. In *Proc. of International Workshop on Video Processing and Quality Metrics for Consumer Electronics*.

[Fetterly et al., 2008] Fetterly, K. A., Blume, H. R., Flynn, M. J., and Samei, E. (2008). Introduction to grayscale calibration and related aspects of medical imaging grade liquid crystal displays. *J. Digit. Imaging*, 21(2):193–207.

[Fiete, 2010] Fiete, R. D. (2010). *Modeling the Imaging Chain of Digital Cameras*. SPIE Press, Bellingham, Washington, USA.

[Firestone et al., 1991] Firestone, L., Cook, K., Culp, K., Talsania, N., and Preston, K. (1991). Comparison of autofocus methods for automated microscopy. *Cytometry*, 12:195–206.

[Frey and Farnand, 2011] Frey, F. and Farnand, S. (2011). Benchmarking art image interchange cycles. Final project report, Rochester Institute of Technology.

[Fryback and Thornbury, 1991] Fryback, D. G. and Thornbury, J. R. (1991). The Efficacy of Diagnostic Imaging. *Medical Decision Making*, 11(2):88–94.

[Fukunaga and Hayes, 1989] Fukunaga, K. and Hayes, R. R. (1989). Effects of sample size in classifier design. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(8):873–885.

[Gallas, 2001] Gallas, B. D. (2001). *Signal detection in lumpy backgrounds*. PhD thesis.

[Gallas, 2006] Gallas, B. D. (2006). One-shot estimate of MRMC variance: AUC. *Acad. Radiol.*, 13 (3):353–62.

[Gallas et al., 2009] Gallas, B. D., Bandos, A., Samuelson, F. W., and Wagner, R. F. (2009). A Framework for Random-Effects ROC Analysis: Biases with the Bootstrap and Other Variance Estimators. *Communications in Statistics-Theory and Methods*, 38(15):2586–2603.

[Gallas and Barrett, 2003] Gallas, B. D. and Barrett, H. H. (2003). Validating the use of channels to estimate the ideal linear observer. *J. Opt. Soc. Am. A*, 20(9):1725–1738.

[Gallas et al., 2013] Gallas, B. D., Cheng, W.-C., Gavrielides, M. A., Keay, T., Wunderlich, A., Hewitt, S. M., Conway, C. M., and Hipp, J. (2013). Evaluating whole slide scanners with task-based reader studies. In *Conference of Medical Image Perception Society (MIPS), Abstracts*, Washington, DC, USA. Medical Image Perception Society (MIPS).

[García-Pérez, 2001] García-Pérez, M. A. (2001). Yes-no staircases with fixed step sizes: psychometric properties and optimal setup. *Optom. Vision Sci.*, 78(1):56–64.

[Gide and Karam, 2012] Gide, M. and Karam, L. (2012). Improved foveation- and saliency-based visual attention prediction under a quality assessment task. In *Proc. IEEE International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 200 –205.

[Gifford et al., 2005] Gifford, H., King, M., Pretorius, P., and Wells, R. (2005). A comparison of human and model observers in multislice LROC studies. *IEEE T. Med. Imaging*, 24(2):160–169.

[Gifford, 2013] Gifford, H. C. (2013). Tests of a 3D visual-search model observer for SPECT. In *Proc. SPIE Medical Imaging*, pages 86730L–86730L–6.

[Gong et al., 2009a] Gong, L., Wang, T., and Liu, F. (2009a). Shape of Gaussians as feature descriptors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2366–2371.

[Gong et al., 2009b] Gong, L., Wang, T., Liu, F., and Chen, G. (2009b). A Lie group based spatiogram similarity measure. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pages 582–585.

[Goossens et al., 2012a] Goossens, B., Luong, Q., Platiša, L., and Philips, W. (2012a). Optimizing image quality using test signals: trading off blur, noise and contrast. In *Proc. IEEE International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 260–265.

[Goossens et al., 2013] Goossens, B., Luong, Q., Platiša, L., and Philips, W. (2013). Objectively measuring signal detectability, contrast, blur and noise in medical images using channelized joint observers. In *Proc. SPIE Medical Imaging*, page 11.

[Goossens et al., 2012b] Goossens, B., Platiša, L., and Philips, W. (2012b). Theoretical performance analysis of multislice channelized Hotelling observers. In Abbey, C. K. and Mello-Thoms, C. R., editors, *Proc. SPIE Medical Imaging*, volume 8318, page 83180U.

[Goossens et al., 2010] Goossens, B., Platiša, L., Vansteenkiste, E., and Philips, W. (2010). The use of steerable channels for detecting asymmetrical signals with random orientations. In Manning, D. J. and Abbey, C. K., editors, *Proc. SPIE Medical Imaging*, volume 7627, page 76270S.

[Green and Swets, 1966] Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley & Sons, New York (reprint Krieger, Melbourne, Florida, 1974).

[Guidelines, 2012] Guidelines (2012). *Guidelines for Quality Control Testing for Digital (CR DR) Mammography v3*. The Royal Australian and New Zealand College of Radiologists.

[Guo and Labate, 2007] Guo, K. and Labate, D. (2007). Optimally sparse multidimensional representation using shearlets. *SIAM Journal on Mathematical Analysis*, 39(1):298–318.

[Hassen et al., 2010] Hassen, R., Wang, Z., and Salama, M. (2010). No-reference image sharpness assessment based on local phase coherence measurement. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2434 –2437.

[He and Park, 2013] He, X. and Park, S. (2013). Model observers in medical imaging research. *Theranostics*, 3(10):774.

[Henricks, 2012] Henricks, W. H. (2012). Evaluation of whole slide imaging for routine surgical pathology: Looking through a broader scope. *J. Pathol. Inform.*, 3(1):39.

[Hillis and Berbaum, 2005] Hillis, S. and Berbaum, K. S. (2005). Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. *Acad. Radiol.*, 12:1534–1542.

[Hillis et al., 2008] Hillis, S., K.S., B., and Metz, C. (2008). Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. *Acad. Radiol.*

[Hillis, 2007] Hillis, S. L. (2007). A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. *Stat. Med.*, 26(3):596–619.

[Hillis et al., 2005] Hillis, S. L., Obuchowski, N. A., Schartz, K. M., and Berbaum, K. (2005). A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette Methods for receiver operating characteristic (ROC) data. *Stat. Med.*, 24:1579–1607.

[Holschneider et al., 1989] Holschneider, M., Kronland-Martinet, R., Morlet, J., and Tchamitchian, P. (1989). Wavelets, time-frequency methods and phase space. *A real-time algorithm for signal analysis with the help of the wavelet transform. Springer, Berlin*, pages 289–297.

[Holst and Lomheim, 2011] Holst, G. C. and Lomheim, T. S. (2011). *CMOS/CCD Sensors and Camera Systems*. JCD Publishing, Winter Park, FL, USA.

[Hsung et al., 1999] Hsung, T., Lun, D., and Siu, W. (1999). Denoising by singularity detection. *IEEE Trans. Signal Proc.*, 47(11):3139–3144.

[Ilić et al., 2009] Ilić, L., Pižurica, A., Vansteenkiste, E., and Philips, W. (2009). Image blur estimation based on the average cone of ratio in the wavelet domain. In Truchetet, F. and Laligant, O., editors, *Proc. SPIE Wavelet Applications in Industrial Processing*, volume 7248, page 72480F.

[ITU-R, 2012] ITU-R (2012). ITU-R Recommendation BT.500-13: Methodology for the Subjective Assessment of the Quality of Television Pictures. Technical report, ITU-R.

[Jaffard, 1991] Jaffard, S. (1991). Pointwise smoothness, two-microlocalization and wavelet coefficients. *Publications Mathematiques*, 35:155–168.

[Jänicke and Chen, 2010] Jänicke, H. and Chen, M. (2010). A salience-based quality metric for visualization. *Computer Graphics Forum*, 29(3):1183–1192.

[Jayaraman et al., 2012] Jayaraman, D., Mittal, A., Moorthy, A., and Bovik, A. (2012). Objective quality assessment of multiply distorted images. In *Proc. Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 1693–1697.

[Jiang et al., 2007] Jiang, Y., Huo, D., and Wilson, D. L. (2007). Methods for quantitative image quality evaluation of MRI parallel reconstructions: detection and perceptual difference model. *Magnetic Resonance Imaging*, 25(5):712 – 721.

[Judy and Swensson, 1985] Judy, P. F. and Swensson, R. G. (1985). Detection of small focal lesions in CT images: effects of reconstruction filters and visual display windows. *Brit. J. Radiol.*, 58:137–145.

[Judy et al., 1981] Judy, P. F., Swensson, R. G., and Szulc, M. (1981). Lesion detection and signal–to–noise ratio in CT images. *Med. Phys.*, 8:13.

[Kanal et al., 2013] Kanal, K., Krupinski, E., Berns, E., Geiser, W., Karellas, A., Mainiero, M., Martin, M., Patel, S., Rubin, D., Shepard, J., Siegel, E., Wolfman, J., Mian, T., Mahoney, M., and Wyatt, M. (2013). ACRAAPMSIIM Practice Guideline for Determinants of Image Quality in Digital Mammography. *J. Digit. Imaging*, 26(1):10–25.

[Kim et al., 2004] Kim, J.-S., Kinahan, P. E., Lartizien, C., Comtat, C., and Lewellen, T. K. (2004). A comparison of planar versus volumetric numerical observers for detection task performance in whole-body PET imaging. *IEEE Trans. Nucl. Sci.*, 51(1):34–40.

[Kimpe and Marchessoux, 2010] Kimpe, T. and Marchessoux, C. (2010). Devices and methods for reducing artefacts in display devices by the use of overdrive.

[Kimpe et al., 2007] Kimpe, T., Marchessoux, C., and Spalla, G. (2007). Evaluating clinical performance of color and grayscale medical displays by mean of a numerical observer. In *Scientific Assembly and Annual Meeting of the Radiological Society of North America*.

[Kimpe et al., 2005] Kimpe, T., Xthona, A., Matthijs, P., and De Paepe, L. (2005). Solution for nonuniformities and spatial noise in medical LCD displays by using pixel-based correction. *J. Digit. Imaging*, 18(3):209–218.

[Koenderink, 1984] Koenderink, J. J. (1984). The structure of images. *Biol Cybern*, 50(5):363–370.

[Krupinski et al., 2013] Krupinski, E. A., Graham, A. R., and Weinstein, R. S. (2013). Characterizing the development of visual search expertise in pathology residents viewing whole slide images. *Human Pathology*, 44(3):357 – 364.

[Krupinski et al., 2012] Krupinski, E. A., Johnson, J. P., Jaw, S., Graham, A. R., and Weinstein, R. S. (2012). Compressing pathology whole-slide images using a human and model observer evaluation. *Journal of Pathology Informatics*, 3.

[Krupinski et al., 2006] Krupinski, E. A., Tillack, A. A., Richter, L., Henderson, J. T., Bhattacharyya, A. K., Scott, K. M., Graham, A. R., Descour, M. R., Davis, J. R., and Weinstein, R. S. (2006). Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. *Human Pathology*, 37(12):1543–1556.

[Kumar et al., 2005] Kumar, S., Biswas, M., and Nguyen, T. (2005). Analysis of the Response Time Compensation system for Liquid Crystal Displays. *European Signal Processing Conference*, 13:.

[Kumcu et al., 2013] Kumcu, A., Elprama, S., Vermeulen, L., Duysburgh, P., Platiša, L., VanNieuwenhove, Y., Van De Winkel, N., Jacobs, A., Van Looy, J., and Philips, W. (2013). Effect of video latency on performance and subjective experience in laparoscopic surgery. In *Conference of Medical Image Perception Society (MIPS), Abstracts*, Washington, DC, USA.

[Kumcu et al., 2012a] Kumcu, A., Platiša, L., Despotovic, I., Vansteenkiste, E., Pizurica, A., Deblaere, K., and Philips, W. (2012a). Multi-modal measurement of cortical thickness in brain MRI for Focal Cortical Dysplasia detection. In *Annual meeting of the International Society for Magnetic Resonance in Medicine (ISMRM) Benelux Chapter, Abstracts*.

[Kumcu et al., 2011a] Kumcu, A., Platiša, L., Goossens, B., and Philips, W. (2011a). Subjective and objective quality evaluation of compressed medical video sequences. In *IEEE International Workshop on Quality of Multimedia Experience (QoMEX), Abstracts*.

[Kumcu et al., 2011b] Kumcu, A., Platiša, L., Goossens, B., and Philips, W. (2011b). Visual quality assessment of H.264 and motion JPEG compressed laparoscopy video. In *IEEE Engineering in Medicine and Biology Society (EMBS) Benelux Chapter, Annual symposium, Abstracts*.

[Kumcu et al., 2011c] Kumcu, A., Platiša, L., Platiša, M., Vansteenkiste, E., Deblaere, K., Badano, A., and Philips, W. (2011c). Trends in reader behavior for a signal detection task in multi- and single-slice images. In *Conference of Medical Image Perception Society (MIPS), Abstracts*.

[Kumcu et al., 2012b] Kumcu, A., Platiša, L., Platiša, M., Vansteenkiste, E., Deblaere, K., Badano, A., and Philips, W. (2012b). Reader behavior in a detection task using single- and multislice image datasets. In Abbey, C. K. and Mello-Thoms, C. R., editors, *Proc. SPIE Medical Imaging*, volume 8318, page 831803.

[Kumcu et al., 2014] Kumcu, A. E., Bombeke, K., Chen, H., Jovanov, L., Platiša, L., Luong, Q., Van Looy, J., Van Nieuwenhove, Y., Schelkens, P., and Philips, W. (2014). Visual quality assessment of H.264/AVC compressed laparoscopy video. In *Proc. SPIE Medical Imaging*. (In press).

[Kupinski et al., 2003] Kupinski, M. A., Hoppin, J. W., Clarkson, E., and Barrett, H. H. (2003). Ideal-observer computation in medical imaging with use of Markov-chain Monte Carlo techniques. *J. Opt. Soc. Am. A*, 20(3):430–438.

[Lagendijk and Biemond, 2009] Lagendijk, R. L. and Biemond, J. (2009). Basic methods for image restoration and identification. The Essential Guide to Image Processing 2nd edition.

[Lange, 2011] Lange, H. (2011). Digital pathology: a regulatory overview. *Lab Medicine*, 42(10):587–591.

[Lartizien et al., 2004] Lartizien, C., Kinahan, P. E., and Comtat, C. (2004). Volumetric model and human observer comparisons of tumor detection for whole-body positron emission tomography. *Acad. Radiol.*, 11(6):637 – 648.

[Lau et al., 2013] Lau, B. A., Das, M., and Gifford, H. C. (2013). Towards visual-search model observers for mass detection in breast tomosynthesis. In *Proc. SPIE Medical Imaging*, pages 86680X–86680X. International Society for Optics and Photonics.

[Leng et al., 2013] Leng, S., Yu, L., Zhang, Y., Carter, R., Toledano, A. Y., and McCollough, C. H. (2013). Correlation between model observer and human observer performance in CT imaging when lesion location is uncertain. *Med. Phys.*, 40(8):081908.

[Levin, 2007] Levin, A. (2007). Blind motion deblurring using image statistics. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 841–848. MIT Press, Cambridge, MA.

[Li et al., 2010] Li, Y., Poulos, A., McLean, D., and Rickard, M. (2010). A review of methods of clinical image quality evaluation in mammography. *Eur. J. Radiol.*, 74(3):e122 – e131.

[Liang and Badano, 2007] Liang, H. and Badano, A. (2007). Temporal response of medical liquid crystal displays. *Med. Phys.*, 34(2):639–646.

[Liang et al., 2008] Liang, H., Park, S., Gallas, B., Myers, K., and Badano, A. (2008). Image browsing in slow medical liquid crystal displays. *Acad. Radiol.*, 15(3):370–82.

[Lin et al., 2004] Lin, J., Zhang, C., and Shi, Q. (2004). Estimating the amount of defocus through a wavelet transform approach. *Pattern Recogn. Lett.*, 25(4):407–411.

[Lindeberg, 1996] Lindeberg, T. (1996). Scale-space: A framework for handling image structures at multiple scales.

[Liu et al., 2013] Liu, H., Engelke, U., Wang, J., Le Callet, P., and Heynderickx, I. (2013). How does image content affect the added value of visual attention in objective image quality assessment? *IEEE Signal Process Lett.*, PP(99):1.

[Liu and Heynderickx, 2011] Liu, H. and Heynderickx, I. (2011). Issues in the Design of a No-Reference Metric for Perceived Blur. In Farnand, SP and Gaykema, F, editor, *Proc. SPIE Electronic Imaging*, volume 7867 of *Proceedings of SPIE*.

[Liu et al., 2012] Liu, H., Koonen, J., Fuderer, M., and Heynderickx, I. (2012). Studying the relative impact of ghosting and noise on the perceived quality of MR images. In *Proc. SPIE Medical Imaging*, pages 83181K–83181K–6.

[Lodge et al., 2009] Lodge, M. A., Rahmim, A., and Wahl, R. L. (2009). A practical, automated quality assurance method for measuring spatial resolution in PET. *J Nucl Med*, 50(8):1307–1314.

[López et al., 2008] López, C., Lejeune, M., Escrivà, P., Bosch, R., Salvadó, M. T., Pons, L. E., Baucells, J., Cugat, X., Álvaro, T., and Jaén, J. (2008). Effects of image compression on automatic count of immunohistochemically stained nuclei in digital images. *Journal of the American Medical Informatics Association*, 15(6):794–798.

[Lukić et al., 2010] Lukić, N., Platiša, L., Pižurica, A., Philips, W., and Temerinac, M. (2010). Real-time wavelet based blur estimation on cell BE platform. In Truchetet, F. and Laligant, O., editors, *Proc. SPIE Electronic Imaging*, volume 7535, page 75350C.

[Luong et al., 2012] Luong, Q., Goossens, B., Aelterman, J., Platiša, L., and Philips, W. (2012). Optimizing image quality in MRI: on the evaluation of k-space trajectories for under-sampled MR acquisition. In *Proc. IEEE International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 25–26.

[Ma et al., 2008] Ma, A. K., Gunn, S., Bullard, E., and Darambara, D. G. (2008). Demonstration of the superiority of digital breast tomosynthesis over 2D mammography through a series of sophisticated computational breast phantoms - a preliminary Monte Carlo study. In *IEEE Nuc. Sci. Symp. Conf. Rec.*, pages 3883 –3885.

[Maidment and Platiša, 2013] Maidment, A. and Platiša, L. (2013). The roles and limitations of model observer studies. In *Conference of Medical Image Perception Society (MIPS), Abstracts*, Washington, DC, USA.

[Malfait and Roose, 1997] Malfait, M. and Roose, D. (1997). Wavelet-based image denoising using a Markov random field a priori model. *IEEE Trans. Image Process.*, 6(4):549 –565.

[Mallat, 1989] Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7):674 – 693.

[Mallat, 1996] Mallat, S. (1996). Wavelets for vision. *Proc. IEEE*, 84(4):604–614.

[Mallat, 1999] Mallat, S. (1999). *A Wavelet Tour of Signal Processing, Second Edition (Wavelet Analysis & Its Applications)*. Academic Press, 2 edition.

[Mallat and Hwang, 1992] Mallat, S. and Hwang, W. (1992). Singularity detection and processing with wavelets. *IEEE Trans on Information Theory*, 38(2, Part 2):617–643.

[Mallat and Zhong, 1992] Mallat, S. and Zhong, S. (1992). Characterization of Signals from Multiscale Edges. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 14(7):710–732.

[Mantiuk et al., 2005] Mantiuk, R., Daly, S., Myszkowski, K., and Seidel, H.-P. (2005). Predicting Visible Differences in High Dynamic Range Images - Model and its Calibration. In Rogowitz, B. E., Pappas, T. N., and Daly, S. J., editors, *Proc. SPIE Human Vision and Electronic Imaging*, volume 5666, pages 204–214.

[Marcelo et al., 2000] Marcelo, A., Fontelo, P., Farolan, M., and Cualing, H. (2000). Effect of image compression on telepathology: a randomized clinical trial. *Archives of pathology & laboratory medicine*, 124(11):1653–1656.

[Marchessoux and Jung, 2006] Marchessoux, C. and Jung, J. (2006). A virtual image chain for perceived image quality of medical display. In *Proc. SPIE Medical Imaging*, pages 61411O–61411O. International Society for Optics and Photonics.

[Marchessoux and Kimpe, 2007] Marchessoux, C. and Kimpe, T. (2007). Specificities of a psychophysical test room dedicated for medical display applications. In *Proc. International symposium, Society for Information Display (SID)*, volume II, pages 971–974, Long Beach.

[Marchessoux et al., 2008a] Marchessoux, C., Kimpe, T., and Bert, T. (2008a). A virtual image chain for perceived and clinical image quality of medical display. *J. Disp. Technol.*, 4:356–368.

[Marchessoux et al., 2008b] Marchessoux, C., Rombaut, A., Kimpe, T., Vermeulen, B., and Demeester, P. (2008b). Extension of a human visual system model for display simulation. In *Proc. SPIE Human Vision and Electronic Imaging*, volume 6806, pages 68061M1–12.

[Marchessoux et al., 2008c] Marchessoux, C., Spalla, G., and Kimpe, T. (2008c). 73.2: A New Methodology for Clinical and Perceived Quality of Medical Displays. In *SID Symposium Digest of Technical Papers*, volume 39, pages 1131–1133. Wiley Online Library.

[Marchessoux et al., 2011] Marchessoux, C., Vivien, N., Kumcu, A., and Kimpe, T. (2011). Validation of a new digital breast tomosynthesis medical display. In *Proc. SPIE Medical Imaging*, volume 7966, page 79660R. SPIE.

[Marr and Nishihara, 1978] Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond., B, Biol. Sci.*, 200(1140):269–294.

[Martens, 2002] Martens, J. (2002). Multidimensional modeling of image quality. *Proc. IEEE*, 90(1):133 –153.

[Marziliano et al., 2004] Marziliano, P., Dufaux, F., Winkler, S., and Ebrahimi, T. (2004). Perceptual blur and ringing metrics: Application to JPEG2000. *Signal Process. Image Commun.*, 19(2):163–172.

[McCartney, 2003] McCartney, R. I. (2003). 48.3: A Liquid Crystal Display Response Time Compensation Feature Integrated into an LCD Panel Timing Controller. *SID Symposium Digest of Technical Papers*, 34(1):1350–1353.

[Mello-Thoms et al., 2011] Mello-Thoms, C., Mello, C. A., Medvedeva, O., Tseytlin, E., and Crowley, R. (2011). Characterizing Virtual Slide Exploration Through the Use of 'Search Maps'. In Manning, D. J. and Abbey, C. K., editor, *Proc. SPIE Medical Imaging*, volume 7966 of *Proceedings of SPIE*. Conference

on Medical Imaging - Image Perception, Observer Performance, and Technology Assessment, Lake Buena Vista, FL, FEB 16-17, 2011.

[Metz, 1993] Metz, C. E. (1993). Quantification of failure to demonstrate statistical significance: the usefulness of confidence intervals. *Investigative radiology*, 28(1):59–63.

[Metz, 2006] Metz, C. E. (2006). Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. *J Am Coll Radiol*, 3(6):413–422.

[Michielsen et al., 2013] Michielsen, K., Zanca, F., Marshall, N., Bosmans, H., and Nuyts, J. (2013). Two complementary model observers to evaluate reconstructions of simulated micro-calcifications in digital breast tomosynthesis. In *Proc. SPIE Medical Imaging*, pages 86730G–86730G. International Society for Optics and Photonics.

[Moorthy and Bovik, 2010] Moorthy, A. K. and Bovik, A. C. (2010). Statistics of natural image distortions. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

[Murthy and Karam, 2010] Murthy, A. and Karam, L. (2010). A MATLAB-based framework for image and video quality evaluation. In *Proc. IEEE International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 242 –247.

[Myers et al., 1986] Myers, K., Barrett, H., Borgstrom, M., Cargill, E., Clough, A., Fiete, R., Milster, T., Patton, D., Paxman, R., Seeley, G., et al. (1986). A systematic approach to the design of diagnostic systems for nuclear medicine. pages 431–444.

[Myers et al., 1985] Myers, K., Barrett, H., Borgstrom, M., Patton, D., and Seeley, G. (1985). Effect of noise correlation on detectability of disk signals in medical imaging. *J. Opt. Soc. Am. A*, 2(10):1752–1759.

[Myers and Barrett, 1987] Myers, K. J. and Barrett, H. H. (1987). Addition of a channel mechanism to the ideal-observer model. *J. Opt. Soc. Am. A*, 4(12):2447–2457.

[Nakamura, 2006] Nakamura, J. (2006). *Image Sensors and Signal Processing for Digital Still Cameras*. CRC Press, Taylor & Francis Group, Boca Raton, FL, USA.

[Narvekar and Karam, 2011] Narvekar, N. and Karam, L. (2011). A No-Reference Image Blur Metric Based on the Cumulative Probability of Blur Detection (CPBD). *IEEE Trans. Image Process.*, 20(9):2678 –2683.

[Narvekar and Karam, 2009] Narvekar, N. D. and Karam, L. J. (2009). A no-reference perceptual image sharpness metric based on a cumulative probability of

blur detection. In *Proc. International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 87–91. International Workshop on Quality of Multimedia Experience (QoMEx 2009), San Diego, CA, JUL 29-31, 2009.

[National Electrical Manufacturers Association (NEMA), 2007] National Electrical Manufacturers Association (NEMA) (2007). *NEMA NU 2-2007 Performance Measurements of Positron Emission Tomographs*. ROSSLYN, Va. The National Electrical Manufacturers Association.

[Nicolosi et al., 2012] Nicolosi, J. S., Yoshida, A. O., Sarian, L. O., Silva, C. A., Andrade, L. A., Derchain, S. F., Vassallo, J., and Schenka, A. A. (2012). Image compression impact on quantitative angiogenesis analysis of ovarian epithelial neoplasms. *Applied Immunohistochemistry & Molecular Morphology*, 20(1):91–95.

[Ninassi et al., 2007] Ninassi, A., Le Meur, O., Le Callet, P., and Barbba, D. (2007). Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. In *Proc. IEEE International Conference on Image Processing (ICIP)*, volume 2, pages II –169 –II –172.

[Obuchowski et al., 2000] Obuchowski, N. A., Lieber, M. L., and Powell, K. A. (2000). Data analysis for detection and localization of multiple abnormalities with application to mammography. *Acad. Radiol.*, 7(7):516–525.

[Oliva and Torralba, 2001] Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *Internat. J. Computer Vision*, 42:145–175.

[Ortiz Jaramillo et al., 2014] Ortiz Jaramillo, B., Kumcu, A., Platia, L., and Philips, W. (2014). A full reference video quality measure based on motion differences and saliency maps evaluation. In *Proc. 9th International Conference on Computer Vision Theory and Applications*, Lisbon, Portugal. (In press).

[Pantanowitz et al., 2013] Pantanowitz, L., Sinard, J. H., Henricks, W. H., Fatheree, L. A., Carter, A. B., Contis, L., Beckwith, B. A., Evans, A. J., Otis, C. N., Lal, A., et al. (2013). Validating Whole Slide Imaging for Diagnostic Purposes in Pathology: Guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Archives of pathology & laboratory medicine*.

[Pantanowitz et al., 2011] Pantanowitz, L., Valenstein, P. N., Evans, A. J., Kaplan, K. J., Pfeifer, J. D., Wilbur, D. C., Collins, L. C., Colgan, T. J., et al. (2011). Review of the current state of whole slide imaging in pathology. *Journal of pathology informatics*, 2(1):36.

[Papp et al., 2009] Papp, I., Lukic, N., Marceta, Z., Teslic, N., and Schu, M. (2009). Real-time video quality assessment platform. In *Proc. International Conference on Consumer Electronics (ICCE)*, pages 1–2.

[Park et al., 2007a] Park, J. M., Franken, E. A., Garg, M., Fajardo, L. L., and Niklason, L. T. (2007a). Breast tomosynthesis: present considerations and future applications. *Radiographics*, 27:S231–S240.

[Park et al., 2009a] Park, S., Badano, A., Gallas, B., and Myers, K. (2009a). Incorporating human contrast sensitivity in model observers for detection tasks. *IEEE Trans. Med. Imaging*, 28(3):339 –347.

[Park et al., 2007b] Park, S., Barrett, H. H., Clarkson, E., Kupinski, M. A., and Myers, K. J. (2007b). Channelized-ideal observer using Laguerre-Gauss channels in detection tasks involving non-Gaussian distributed lumpy backgrounds and a Gaussian signal. *J. Opt. Soc. Am. A*, 24(12):B136–B150.

[Park and Clarkson, 2009] Park, S. and Clarkson, E. (2009). Efficient estimation of ideal-observer performance in classification tasks involving high-dimensional complex backgrounds. *J. Opt. Soc. Am. A*, 26(11):B59.

[Park et al., 2006] Park, S., Clarkson, E., Barrett, H. H., Kupinski, M. A., and Myers, K. J. (2006). Performance of a channelized-ideal observer using laguerre-gauss channels for detecting a gaussian signal at a known location in different lumpy backgrounds.

[Park et al., 2005] Park, S., Clarkson, E., Kupinski, M. A., and Barrett, H. H. (2005). Efficiency of the human observer detecting random signals in random backgrounds. *J. Opt. Soc. Am. A*, 22(1):3–16.

[Park et al., 2007c] Park, S., Gallas, B. D., Badano, A., Petrick, N. A., and Myers, K. J. (2007c). Efficiency of the human observer for detecting a Gaussian signal at a known location in non-Gaussian distributed lumpy backgrounds. *J. Opt. Soc. Am. A*, 24(4):911–921.

[Park et al., 2010] Park, S., Jennings, R., Liu, H., Badano, A., and Myers, K. (2010). A statistical, task-based evaluation method for three-dimensional x-ray breast imaging systems using variable-background phantoms. *Med. Phys.*, 37:6253 – 6270.

[Park et al., 2003] Park, S., Kupinski, M., Clarkson, E., and Barrett, H. (2003). *Ideal-observer performance under signal and background uncertainty*, volume 2732, chapter Inf Process Med Imaging, pages 342–353.

[Park et al., 2009b] Park, S., Witten, J. M., and Myers, K. J. (2009b). Singular vectors of a linear imaging system as efficient channels for the Bayesian ideal observer. *IEEE T. Med. Imaging*, 28(5):657–668.

[Parwani et al., 2011] Parwani, A., Chubb, L., Pantanowitz, L., and Singh, R. (2011). Standardization in digital pathology: Supplement 145 of the DICOM standards. Technical Report 1.

[Pelli, 1985] Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *J. Opt. Soc. Am. A*, 2(9):1508–1531.

[Philp et al., 2013] Philp, J. C., Frieden, I. J., and Cordoro, K. M. (2013). Pediatric teledermatology consultations: Relationship between provided data and diagnosis. *Pediatric dermatology*, 30(5):561–567.

[Pinson and Wolf, 2003] Pinson, M. H. and Wolf, S. (2003). Comparing subjective video quality testing methodologies. In *Visual Communications and Image Processing 2003*, pages 573–582. International Society for Optics and Photonics.

[Pižurica, 2002] Pižurica, A. (2002). *Image denoising using wavelets and spatial context modeling*. PhD thesis, Ghent University.

[Pižurica et al., 2002] Pižurica, A., Philips, W., Lemahieu, I., and Acheroy, M. (2002). A joint inter- and intrascale statistical model for Bayesian wavelet based image denoising. *IEEE Trans. Image Process.*, 11(5):545–557.

[Pižurica et al., 2013] Pižurica, A., Platiša, L., Ružić, T., Cornelis, B., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2013). *Het Lam Gods Series of Lectures*, chapter Virtual Restoration and Mathematical Analysis of Pearls in the Adoration of the Mystic Lamb. (To appear).

[Platiša, 2008] Platiša, L. (2008). Optimization of medical imaging display systems using the channelized Hotelling observer. In *UGent-FirW Doctoraatssymposium, Abstracts*, pages 194–195.

[Platiša, 2010] Platiša, L. (2010). No-reference wavelet-based blur metric for image quality assessment. In *UGent-FirW Doctoraatssymposium, Abstracts*, pages 152–152. Universiteit Gent. Faculteit Ingenieurswetenschappen.

[Platiša et al., 2010a] Platiša, L., Cornelis, B., Ružić, T., Pižurica, A., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2010a). Pearls and beads in Jan van Eyck's paintings. In *Vision and material : interaction between art and science in Jan Van Eyck's time, Abstracts*.

[Platiša et al., 2011a] Platiša, L., Cornelis, B., Ružić, T., Pižurica, A., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2011a). Spatiogram features to characterize pearls in paintings. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 801–804.

[Platiša et al., 2012a] Platiša, L., Cornelis, B., Ružić, T., Pižurica, A., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2012a). Spatiogram features to characterize pearls in the Ghent Altarpiece. In *Symposium for the Study of Underdrawing and Technology in Painting, Abstracts*. Ghent University, Department of Telecommunications and information processing.

[Platiša et al., 2012b] Platiša, L., Cornelis, B., Ružić, T., Pižurica, A., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2012b). *Vision and material : interaction between art and science in Jan Van Eyck's time*, chapter Spatiogram Features to Characterize Pearls and Beads and other Small Ball-shaped Objects in Paintings, pages 315–329. KVAB PRESS.

[Platiša et al., 2014a] Platiša, L., Cornelis, B., Ružić, T., Pižurica, A., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2014a). Spatiogram-based descriptors for quality of appearance of pearl-like objects in the images. (In preparation).

[Platiša et al., 2011b] Platiša, L., Cornelis, B., Ružić, T., Pižurica, A., Dooms, A., Schelkens, P., Mar-tens, M., De Mey, M., and Daubechies, I. (2011b). Spatiogram features to characterize pearls in paintings. In *International workshop on Image Processing for Art Investigation, Abstracts*.

[Platiša et al., 2011c] Platiša, L., De Smet, A., Despotović, I., Kumcu, A., Vansteenkiste, E., Deblaere, K., Pižurica, A., and Philips, W. (2011c). Measuring cortical thickness in brain MRI volumes to detect focal cortical dysplasia (FCD) in epilepsy patients. In *Annual meeting of the International Society for Magnetic Resonance in Medicine (ISMRM) Benelux Chapter, Abstracts*.

[Platiša et al., 2011d] Platiša, L., De Smet, A., Despotović, I., Kumcu, A., Vansteenkiste, E., Deblaere, K., Pizurica, A., and Philips, W. (2011d). Measuring cortical thickness in brain MRI volumes to detect focal cortical dysplasia (FCD) in epilepsy patients. In *Proc. International Society for Magnetic Resonance in Medicine (ISMRM)*, number 19, pages 2438–2438.

[Platiša et al., 2009a] Platiša, L., Goossens, B., Vansteenkiste, E., Badano, A., and Philips, W. (2009a). Channelized Hotelling observers for detection tasks in multislice images. In *Conference of Medical Image Perception Society (MIPS), Abstracts*, pages 36–36.

[Platiša et al., 2009b] Platiša, L., Goossens, B., Vansteenkiste, E., Badano, A., and Philips, W. (2009b). Channelized Hotelling observers for the detection of 2D signals in 3D simulated images. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 1761–1764.

[Platiša et al., 2010b] Platiša, L., Goossens, B., Vansteenkiste, E., Badano, A., and Philips, W. (2010b). Quantifying temporal effects of medical LCD monitors on lesion detectability. In *Belgian Day on Biomedical Engineering, 9th, Abstracts*.

[Platiša et al., 2010c] Platiša, L., Goossens, B., Vansteenkiste, E., Badano, A., and Philips, W. (2010c). Using channelized Hotelling observers to quantify temporal effects of medical liquid crystal displays on detection performance. In Manning, D. J. and Abbey, C. K., editors, *Proc. SPIE Medical Imaging*, volume 7627, page 76270U.

[Platiša et al., 2011e] Platiša, L., Goossens, B., Vansteenkiste, E., Park, S., Gallas, B. D., Badano, A., and Philips, W. (2011e). Channelized Hotelling observers for the assessment of volumetric imaging data sets. *J. Opt. Soc. Am. A*, 28(6):1145–1163.

[Platiša et al., 2012c] Platiša, L., Kumcu, A., Platiša, M., Vansteenkiste, E., Deblaere, K., Badano, A., and Philips, W. (2012c). Volumetric detection tasks with varying complexity: human observer performance. In Abbey, C. K. and Mello-Thoms, C. R., editors, *Proc. SPIE Medical Imaging*, volume 8318, page 83180S.

[Platiša et al., 2011f] Platiša, L., Kumcu, A., Platiša, M., Vansteenkiste, E., Gallas, B., Deblaere, K., Badano, A., and Philips, W. (2011f). Model and human observers studies in volumetric images for detection tasks with varying complexity. In *Conference of Medical Image Perception Society (MIPS), Abstracts*, pages 27–27.

[Platiša et al., 2014b] Platiša, L., Kumcu, A., Platiša, M., Vansteenkiste, E., Gallas, B. D., Deblaere, K., Badano, A., and Philips, W. (2014b). Lesion detection performance in single- versus multi-slice image readings: results from human and model observer studies. (In preparation).

[Platiša et al., 2010d] Platiša, L., Lukić, N., Pižurica, A., Vansteenkiste, E., and Philips, W. (2010d). Image blur estimation based on the average cone of ratio in the wavelet domain. In *Sparsity and modern mathematical methods for high dimensional data, Abstracts*, pages 24–24.

[Platiša et al., 2011g] Platiša, L., Marchessoux, C., Goossens, B., and Philips, W. (2011g). Performance evaluation of medical LCD displays using 3D channelized Hotelling observers. In Manning, D. J. and Abbey, C. K., editors, *Proc. SPIE Medical Imaging*, volume 7966, page 79660T.

[Platiša et al., 2011h] Platiša, L., Marchessoux, C., Kimpe, T., Vansteenkiste, E., Badano, A., and Philips, W. (2011h). Channelized Hotelling observers for signal detection in stack-mode reading of volumetric images on medical displays with slow response time. In *Proc. IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pages 2697–2702.

[Platiša and Pižurica, 2014] Platiša, L. and Pižurica, A. (2014). No-reference blur estimation based on the average cone ratio in the wavelet domain. (In preparation).

[Platiša et al., 2009c] Platiša, L., Pižurica, A., Vansteenkiste, E., and Philips, W. (2009c). Image blur estimation based on the average cone of ratio in the wavelet domain. In Truchetet, F. and Laligant, O., editors, *Proc. SPIE Electronic Imaging*, volume 7248, page 10.

[Platiša et al., 2011i] Platiša, L., Pižurica, A., Vansteenkiste, E., and Philips, W. (2011i). No-reference blur estimation based on the average cone ratio in the wavelet

domain. In *IEEE International Workshop on Quality of Multimedia Experience (QoMEX), Abstracts*.

[Platiša et al., 2011j] Platiša, L., Pizurica, A., Vansteenkiste, E., and Philips, W. (2011j). No-reference blur estimation based on the average cone ratio in the wavelet domain. In Akopian, D., Creutzburg, R., Snoek, C. G. M., Sebe, N., and Kennedy, L., editors, *Proc. SPIE Electronic Imaging*, volume 7881, page 78811D.

[Platiša et al., 2013a] Platiša, L., Van Brantegem, L., Kumcu, A., Marchessoux, C., Vansteenkiste, E., and Philips, W. (2013a). Effects of common image manipulations on diagnostic performance in digital pathology human study. In *Conference of Medical Image Perception Society (MIPS), Abstracts*, Washington, DC, USA.

[Platiša et al., 2013b] Platiša, L., Van Brantegem, L., Vander Haeghen, Y., Marchessoux, C., Vansteen-kiste, E., and Philips, W. (2013b). Psycho-visual evaluation of image quality attributes in digital pathology slides viewed on a medical color LCD display. In *Proc. SPIE Medical Imaging*, volume 8676, page 86760J, Orlando, Florida, USA.

[Platiša et al., 2009d] Platiša, L., Vansteenkiste, E., Goossens, B., Marchessoux, C., Kimpe, T., and Philips, W. (2009d). Optimization of medical imaging display systems: using the channelized Hotelling observer for detecting lung nodules: experimental study. In Sahiner, B. and Manning, D. J., editors, *Proc. SPIE Medical Imaging*, volume 7263, page 72630P.

[Polatkan et al., 2009] Polatkan, G., Jafarpour, S., Brasoveanu, A., Hughes, S. M., and Daubechies, I. (2009). Detection of forgery in paintings using supervised learning. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 2921–2924. IEEE.

[Popescu and Myers, 2013] Popescu, L. M. and Myers, K. J. (2013). CT image assessment by low contrast signal detectability evaluation with unknown signal. *Med. Phys.*, 40(11):111908.

[Qu et al., 2013] Qu, X., Kumcu, A., Platiša, L., Despotovic, I., Deblaere, K., Bai, T., and Philips, W. (2013). Blur estimation at the gray-white matter boundary for focal cortical dysplasia in magnetic resonance imaging. In *IEEE Annual International Conference of the Engineering in Medicine and Biology Society (EMBS), Abstracts*.

[Qu et al., 2014] Qu, X., Platiša, L., Despotovic, I., Kumcu, A., Deblaere, K., Bai, T., and Philips, W. (2014). Automatic brain atlas in magnetic resonance image for focal cortical dysplasia patients. In *IEEE International Symposium on Biomedical Imaging (ISBI), Abstracts*. (In press).

[Rafferty et al., 2013] Rafferty, E. A., Park, J. M., Philpotts, L. E., Poplack, S. P., Sumkin, J. H., Halpern, E. F., and Niklason, L. T. (2013). Assessing radiologist performance using combined digital mammography and breast tomosynthesis

compared with digital mammography alone: results of a multicenter, multireader trial. *Radiol.*, 266(1):104–113.

[Rahmim and Zaidi, 2008] Rahmim, A. and Zaidi, H. (2008). PET versus SPECT: strengths, limitations and challenges. *Nuclear Medicine Communications*, 29:193–207.

[Redi et al., 2010] Redi, J., Liu, H., Alers, H., Zunino, R., and Heynderickx, I. (2010). Comparing Subjective Image Quality Measurement Methods for the Creation of Public Databases. In Farnand, SP and Gaykema, F, editor, *Proc. SPIE Image Quality and System Performance*, volume 7529 of *Proceedings of SPIE*. Conference on Image Quality and System Performance VII, San Jose, CA, JAN 18-19, 2010.

[Redondo et al., 2012] Redondo, R., Bueno, G., Cristóbal, G., Vidal, J., Déniz, O., García-Rojo, M., Murillo, C., Relea, F., and González, J. (2012). Quality evaluation of microscopy and scanned histological images for diagnostic purposes. *Micron*, 43(2):334–343.

[Reiner, 2013] Reiner, B. I. (2013). Creating Accountability in Image Quality Analysis. Part 2: Medical Imaging Accreditation. *Journal of digital imaging*, pages 1–4.

[Reiner et al., 2001] Reiner, B. I., Siegel, E. L., Hooper, F. J., Pomerantz, S., Dahlke, A., and Rallis, D. (2001). Radiologists' productivity in the interpretation of CT scans: a comparison of PACS with conventional film. *Am. J. Roentgenol.*, 176:861–864.

[Rojo et al., 2006] Rojo, M. G., García, G. B., Mateos, C. P., García, J. G., and Vicente, M. C. (2006). Critical comparison of 31 commercially available digital slide systems in pathology. *International Journal of Surgical Pathology*, 14(4):285–305.

[Rolland and Barrett, 1992] Rolland, J. P. and Barrett, H. H. (1992). Effect of random background inhomogeneity on observer detection performance. *J. Opt. Soc. Am. A*, 9(5):649–658.

[Romeny, 1996] Romeny, B. M. T. H. (1996). Introduction to scale-space theory: Multiscale geometric image analysis. Technical report, Verlag. First International Conference on Scale-Space theory.

[Rooms et al., 2002] Rooms, F., Pizurica, A., and Philips, W. (2002). Estimating image blur in the wavelet domain. In Suter, D. and Bab-Hadiashar, A., editors, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages IV–4190.

[Rosenfeld and Thurston, 1971] Rosenfeld, A. and Thurston, M. (1971). Edge and curve detection for visual scene analysis. *IEEE Trans. Comput.*, C-20(5):562–569.

[Rousson et al., 2014] Rousson, J., Couturou, J., Vetsuypens, A., Platiša, L., Kumcu, A., Kimpe, T., and Philips, W. (2014). Subjective quality and depth assessment in stereoscopic viewing of volume-rendered medical images. In *Proc. SPIE Electronic Imaging*. (In press).

[Ružić et al., 2010] Ružić, T., Cornelis, B., Platiša, L., Pižurica, A., Dooms, A., Martens, M., De Mey, M., and Daubechies, I. (2010). Craquelure inpainting in art work. In *Vision and material : interaction between art and science in Jan Van Eyck's time, Abstracts*.

[Ružić et al., 2011] Ružić, T., Cornelis, B., Platiša, L., Pižurica, A., Dooms, A., Philips, W., Martens, M., De Mey, M., and Daubechies, I. (2011). Virtual restoration of the Ghent altarpiece using crack detection and inpainting. In Blanc-Talon, J., Kleihorst, R., Philips, W., Popescu, D., and Scheunders, P., editors, *Lecture Notes in Computer Science*, volume 6915, pages 417–428.

[Sadler and Swami, 1999] Sadler, B. M. and Swami, A. (1999). Analysis of multiscale products for step detection and estimation. *IEEE Trans. Inf. Theory*, 45(3):1043–1051.

[Samei et al., 2005] Samei, E., Badano, A., Chakraborty, D., Compton, K., Cornelius, C., Corrigan, K., Flynn, M. J., Hemminger, B., Hangiandreou, N., Johnson, J., et al. (2005). Assessment of display performance for medical imaging systems: executive summary of AAPM TG18 report. *Med. Phys.*, 32:1205.

[Samei et al., 1997] Samei, E., Flynn, M., and Eyler, W. (1997). Simulation of subtle lung nodules in projection chest radiography. *Radiol.*, 202(1):117–124.

[Samei et al., 2003] Samei, E., Flynn, M., Peterson, E., and Eyler, W. (2003). Subtle lung nodules: Influence of local anatomic variations on detection. *Radiol.*, 228(1):76–84.

[Samei et al., 1999] Samei, E., Flynn, M. J., and Eyler, W. R. (1999). Detection of subtle lung nodules: Relative influence of quantum and anatomic noise on chest radiographs. *Radiol.*, 213(3):727–734.

[Seidenari et al., 2004] Seidenari, S., Pellacani, G., Righi, E., and Nardo, A. D. (2004). Is JPEG compression of videomicroscopic images compatible with telediagnosis? Comparison between diagnostic performance and pattern recognition on uncompressed TIFF images and JPEG compressed ones. *Telemedicine Journal & e-Health*, 10(3):294–303.

[Sheikh et al., 2006] Sheikh, H., Sabir, M., and Bovik, A. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 15(11):3440 –3451.

[Shidahara et al., 2006] Shidahara, M., Inoue, K., Maruyama, M., Watabe, H., Taki, Y., Goto, R., Okada, K., Khmomura, S., Osawa, S., Onishi, Y., et al. (2006). Predicting human performance by channelized Hotelling observer in discriminating between Alzheimers dementia and controls using statistically processed brain perfusion SPECT. *Annals of nuclear medicine*, 20(9):605–613.

[Shiraishi et al., 2000] Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.-i., Matsui, M., Fujita, H., Kodera, Y., and Doi, K. (2000). Development of a digital image database for chest radiographs with and without a lung nodule receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *Am. J. Roentgenol.*, 174(1):71–74.

[Silver, 2006] Silver, L. (2006). Arts and Minds: Scholarship on Early Modern Art History (Northern Europe). *Renaissance Quarterly*, 59(2):pp. 351–373.

[Sinisgalli, 2006] Sinisgalli, R. (2006). *Alberti's "De Pictura". Il Nuovo De Pictura di Leon Battista Alberti, The New De Pictura of Leon Battista Alberti*. Edizioni Kappa, Roma.

[Soleimani et al., 2013] Soleimani, S., Rooms, F., and Philips, W. (2013). Efficient blur estimation using multi-scale quadrature filters. *Signal Process.*, 93(7):1988 – 2002.

[Starr et al., 1975] Starr, S. J., Metz, C. E., Lusted, L. B., and Goodenough, D. J. (1975). Visual detection and localization of radiographic images. *Radiol.*, 116(3):533–538.

[Stork, 2006] Stork, D. (2006). Computer Vision, Image Analysis, and Master Art: Part 1. *IEEE Multimedia*, 13(3):16–20.

[Stork and Duarte, 2007] Stork, D. and Duarte, M. (2007). Computer Vision, Image Analysis, and Master Art: Part 3. *IEEE Multimedia*, 14(1):14–18.

[Stork and Johnson, 2006] Stork, D. and Johnson, M. (2006). Computer Vision, Image Analysis, and Master Art: Part 2. *IEEE Multimedia*, 13(4):12–17.

[Stromeyer and Julesz, 1972] Stromeyer, C. F. I. and Julesz, B. (1972). Spatial-frequency masking in vision: Critical bands and spread of masking. *J. Opt. Soc. Am. A*, 62(10):1221–1232.

[Swets and Pickett, 1982] Swets, J. and Pickett, R. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*.

[Tabár et al., 2011] Tabár, L., Vitak, B., Chen, T. H.-H., Yen, A. M.-F., Cohen, A., Tot, T., Chiu, S. Y.-H., Chen, S. L.-S., Fann, J. C.-Y., Rosell, J., et al. (2011). Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiol.*, 260(3):658–663.

[Taplin et al., 2002] Taplin, S. H., Rutter, C. M., Finder, C., Mandelson, M. T., Houn, F., and White, E. (2002). Screening mammography: clinical image quality and the risk of interval breast cancer. *Am. J. Roentgenol.*, 178(4):797–803.

[Tong et al., 2004] Tong, H., Li, M., Zhang, H., and Zhang, C. (2004). Blur detection for digital images using wavelet transform. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, volume 1, pages 27–30.

[Ulges et al., 2006] Ulges, A., Lampert, C. H., and Keysers, D. (2006). Spatiogram-based shot distances for video retrieval. In *TREC Video Retrieval Evaluation (TRECVID) Workshop*. Citeseer.

[Van den Branden Lambrecht, 1996] Van den Branden Lambrecht, C. J. (1996). A working spatio-temporal model of the human visual system for image restoration and quality assessment applications. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 2291–2294 vol. 4.

[Vansteenkiste et al., 2006] Vansteenkiste, E., Van der Weken, D., Philips, W., and Kerre, E. (2006). Evaluation of the perceptual performance of fuzzy image quality measures. In Gabrys, B., Howlett, R., and Jain, L., editors, *Lecture Notes in Computer Science*, volume 4251, pages 623–630. Springer Berlin, Heidelberg.

[Vaz et al., 2011] Vaz, M., Besnehard, Q., and Marchessoux, C. (2011). 3D Lesions Insertion in Digital Breast Tomosynthesis Images. In *Proc. SPIE Medical Imaging*, volume 7961, page 79615Z.

[Verougstraete et al., 2004] Verougstraete, H., Van Schoute, R., and Borchert, T. H. (2004). *Fake or not fake. Het verhaal van de restauratie van de Vlaamse Primitieven*. Ludion, Ghent-Amsterdam.

[Vetsuypens et al., 2011] Vetsuypens, A., Besnehard, Q., Platiša, L., Arnault, E., Marchessoux, C., Kimpe, T., Xthona, A., and Philips, W. (2011). Creating a modality-optimized medical display for DBT based on MEVIC simulations. In *Conference of Medical Image Perception Society (MIPS), Abstracts*.

[Vladár et al., 1998] Vladár, A. E., Postek, M. T., Davidson, M., et al. (1998). Image sharpness measurement in scanning electron microscopypart II. *Scanning*, 20(1):24–34.

[Vu et al., 2012] Vu, C., Phan, T., and Chandler, D. (2012). S3 : A spectral and spatial measure of local perceived sharpness in natural images. *IEEE Trans. Image Process.*, 21(3):934–945.

[Vu and Chandler, 2012] Vu, P. and Chandler, D. (2012). A fast wavelet-based algorithm for global and local image sharpness estimation. *IEEE Signal Process Lett.*, 19(7):423 –426.

[Wang et al., 2004a] Wang, H., Wu, T., Zhu, X., and Wu, S. (2004a). Correlations between liquid crystal director reorientation and optical responsetime of a homeotropic cell. *J. Appl. Phys.*, 95(10):5502–5508.

[Wang, 2011] Wang, Z. (2011). Applications of objective image quality assessment methods [applications corner]. *IEEE Signal Process. Mag.*, 28(6):137 –142.

[Wang et al., 2004b] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004b). Image quality assessment: from error visibility to structural similarity. *IEEE T. Med. Imaging*, 13(4):600 –612.

[Wang and Bovik, 2006] Wang, Z. and Bovik, A. C. (2006). *Modern Image Quality Assessment*. Synthesis Lectures on Image, Video, and Multimedia Processing. Morgan & Claypool Publishers.

[Watson and Kramer, 1999] Watson, S. E. and Kramer, A. F. (1999). Object-based visual selective attention and perceptual organization. *Perception & Psychophysics*, 61(1):31–49.

[Weinstein et al., 2009] Weinstein, R. S., Graham, A. R., Richter, L. C., Barker, G. P., Krupinski, E. A., Lopez, A. M., Erps, K. A., Bhattacharyya, A. K., Yagi, Y., and Gilbertson, J. R. (2009). Overview of telepathology, virtual microscopy, and whole slide imaging: prospects for the future. *Human Pathology*, 40(8):1057–1069.

[Wells et al., 2000] Wells, R., King, M., Gifford, H., and Pretorius, P. (2000). Single-slice versus multi-slice display for human-observer lesion-detection studies. *IEEE Trans. Nucl. Sci.*, 47(3):1037–1044.

[Williams et al., 2007] Williams, M. B., Krupinski, E. A., Strauss, K. J., Breeden III, W. K., Rzeszotarski, M. S., Applegate, K., Wyatt, M., Bjork, S., and Seibert, J. A. (2007). Digital radiography image quality: image acquisition. *J. Am. Coll. Radiol.*, 4(6):371–388.

[Winkler, 2009] Winkler, S. (2009). On the properties of subjective ratings in video quality experiments. In *Proc. IEEE International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 139–144. IEEE.

[Winkler, 2012] Winkler, S. (2012). Analysis of public image and video databases for quality assessment. *EEE J. Sel. Top. Sign. Proces.*, PP(99):1.

[Witkin and Tenenbaum, 1983] Witkin, A. P. and Tenenbaum, J. M. (1983). *Human and Machine Vision*, chapter On the role of structure in vision, pages 481–543. Academic Press, New York.

[Witten et al., 2009] Witten, J. M., Park, S., and Myers, K. J. (2009). Using partial least squares to compute efficient channels for the Bayesian ideal observer. In Sahiner, B. and Manning, D. J., editors, *Proc. SPIE Medical Imaging*, volume 7263, page 72630Q. SPIE.

[Wolfe, 2013] Wolfe, J. M. (2013). When is it time to move to the next raspberry bush? Foraging rules in human visual search. *J. Vis.*, 13(3):10.

[Yagi and Gilbertson, 2005] Yagi, Y. and Gilbertson, J. (2005). Digital imaging in pathology: the case for standardization. *J. Telemed. Telecare*, 11(3):109–116.

[Yagi and Pantanowitz, 2012] Yagi, Y. and Pantanowitz, L. (2012). Comment on "Quality evaluation of microscopy and scanned histological images for diagnostic purposes" : Are scanners better than microscopes? *J. Pathol. Inform.*, 3:14.

[Yao et al., 2006] Yao, Y., Abidi, B., Doggaz, N., and Abidi, M. (2006). Evaluation of sharpness measures and search algorithms for the auto focusing of high-magnification images. In *Proc. SPIE Conference on Visual Information Processing*, pages 62460G–62460G–12.

[Yao et al., 2011] Yao, Z., Lai, Z., and Liu, W. (2011). A symmetric KL divergence based spatiogram similarity measure. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 193–196.

[Yoshida, 2005] Yoshida, H. (2005). *Image Sensors and Signal Processing for Digital Still Cameras*, chapter Evaluation of Image Quality, pages 277–304. CRC Press.

[Young et al., 2010] Young, K. C., Van Engen, R., Bosmans, H., Jacobs, J., and Zanca, F. (2010). Quality control in digital mammography. In *Digital Mammography*, pages 33–54. Springer.

[Young et al., 2009] Young, S., Park, S., Anderson, S. K., Badano, A., Myers, K. J., and Bakic, P. (2009). Estimating breast tomosynthesis performance in detection tasks with variable-background phantoms. In Samei, E. and Hsieh, J., editors, *Proc. SPIE Medical Imaging*, volume 7258, page 72580O. SPIE.

[Zanca et al., 2012] Zanca, F., Van Ongeval, C., Claus, F., Jacobs, J., Oyen, R., and Bosmans, H. (2012). Comparison of visual grading and free-response ROC analyses for assessment of image processing algorithms in digital mammography. *Brit. J. Radiol.*

[Zhan and Karam, 2003] Zhan, C. and Karam, L. (2003). Wavelet-based adaptive image denoising with edge preservation. In *Proc. IEEE International Conference on Image Processing (ICIP)*, volume 1, pages I–97–100 vol.1.

[Zhang and Bao, 2002] Zhang, L. and Bao, P. (2002). Edge detection by scale multiplication in wavelet domain. *Pattern Recognit. Lett.*, 23(14):1771–1784.

[Zhang et al., 2012] Zhang, L., Cavaro-Menard, C., Le Callet, P., and Tanguy, J.-Y. (2012). A Perceptually relevant Channelized Joint Observer (PCJO) for the detection-localization of parametric signals. *IEEE T. Med. Imaging*, 31(10):1875 –1888.

[Zhang et al., 2013] Zhang, L., Goossens, B., Cavaro-Ménard, C., Callet, P. L., and Ge, D. (2013). Channelized model observer for the detection and estimation of signals with unknown amplitude, orientation, and size. *J. Opt. Soc. Am. A*, 30(11):2422–2432.

[Zhang et al., 1997] Zhang, N., Postek, M., Larrabee, R., Vladar, A., Keery, W., and Jones, S. (1997). A statistical measure for the sharpness of the SEM images. In *Proc. SPIE*, volume 3050.

[Zhang et al., 2005] Zhang, N., Vladar, A., Postek, M., and Larrabee, R. (2005). Spectral density-based statistical measures for image sharpness. *Metrologia*, 42(5):351–359.

[Zhang et al., 2003] Zhang, N., Vladar, A. E., Postek, M. T., and Larrabee, B. (2003). A kurtosis-based statistical measure for two-dimensional processes and its application to image sharpness. In *Proc. Section of Physical and Engineering Sciences of American Statistical Society*.

[Zhang et al., 2007] Zhang, Y., Pham, B. T., and Eckstein, M. P. (2007). Evaluation of internal noise methods for Hotelling observer models. *Med. Phys.*, 34(8):3312–3322.

[Zhou et al., 2011] Zhou, X.-H., Obuchowski, N. A., and McClish, D. K. (2011). *Statistical methods in diagnostic medicine*, volume 712. Wiley Series in Probability and Statistics.